Abstract—Fairness, in general, is a topic that has received much attention in research on distributed systems. In their application as electronic trading venues, however, temporal fairness remains a topic that is poorly understood. This is concerning because operators of these venues generally have obligations to ensure their fairness. Consequently, this paper (1) describes what temporal fairness is and is not, (2) identifies things that can make it elusive, and (3) describes a mechanism for improving it that was recently retrofitted to a major FX trading venue: Thomson Reuters Matching.

I. INTRODUCTION

Electronic Trading Venues (ETVs) are an important and widely-used application of distributed systems. Real-life examples of such systems include the New York Stock Exchange, the Chicago Mercantile Exchange, Thomson Reuters Matching (TRM) and so on. ETVs are important because in aggregate, world-wide they facilitate the daily exchange of literally trillions of dollars worth of financial instruments [1].

In recent years there has been no shortage of controversy—even in mainstream media—about fairness pertaining to speed on ETVs, or concerning the perceived lack thereof [2]–[7]. What is interesting about this controversy is that it has persisted over several years (from 2012 to present), it spans multiple jurisdictions (New York, Chicago, London, India, Korea) and multiple asset classes (futures, foreign exchange and equities). In a recent talk on his ongoing work in this area an academic economist implied issues of temporal unfairness could, in the US stock market alone, be causing the misappropriation of $200 billion per year away from the pockets of everyday investors [8]. In all senses the problem then, even if it were only one of perception, does seem to be quite serious and quite widespread.

What would seem be helpful in distinguishing the actual from the perceived is a more careful description of what the problem is (and is not). Without such a description it has been noted arguments pertaining to fairness can easily become emotive, political, philosophical and so on [9]. A careful description of the problem is further necessary because operators of ETVs generally have obligations to ensure the markets those venues host are fair and orderly [10]–[12].

Operators of those ETVs would like to be able to prove (e.g., to regulators) that they are meeting those obligations [13] [14].

The contributions of this paper, by section, are as follows. Section II provides a careful description of what the problem of temporal fairness is and is not, identifies specific reasons why notions of it appearing in securities laws are impractical to implement on ETVs, and consequently proposes a practicable, stochastic definition of temporal fairness. Section III describes the mechanism implemented by TRM for improving temporal fairness and the criteria that led to its selection over other such mechanisms. Section IV, besides concluding this work, identifies as future work the need to formalize equivalence between the same exact time form of temporal fairness with the stochastic form of it, and questions the need for an even stronger form of the latter.

II. THE MEANING OF TEMPORAL FAIRNESS

There are quite a large number of securities laws in existence that specify information must be disseminated to market participants at the same exact time [15]. These laws exist to ensure a fair allocation of scarce resources among market participants. Insider trading, for instance, is prohibited because it has been deemed unfair to allow a person with non-public information about a company (e.g., that in this quarter the company has far exceeded its sales forecasts) to buy its stock at a low price ahead of all other investors, who as a result of this information ultimately becoming public, will likely have to pay a higher price for it. By virtue of the information being non-public the insider trader has received it and acted upon it before other investors; the scarce resource here is the lower-priced stock (and not the stock in general).

The problem with this requirement for same exact time dissemination of information is that what are perhaps quite subtle characteristics of the technology used to implement ETVs likely makes it impossible to meet. What is seemingly possible though, given these technology constraints, is to ensure that over a long time horizon where there are continual, distinct and ongoing disclosures of information, that the allocation of scarce resources ends up the same as if at each instance of information disclosure every participant was sent it at the same exact time.
The other issue, as alluded to earlier, is distinguishing what might only be a perceived problem of temporal fairness from that which is an actual problem. In short the actual problem manifests as varying transmission and processing times among the constituent components of the ETV (by virtue of them being constituent, they are typically owned by or under the control of the ETV’s operator). The perceived problem—which typically does not feature in the body of law in this area\(^1\), and as we shall see shortly is perhaps better described as a social welfare issue—pertains to transmission and processing times in components owned by or under the control of individual market participants.

A. Closed- and Open-Loop Forms of the Problem

There are two forms of the problem of temporal fairness as it manifests on an ETV: closed-loop and open-loop. These forms are illustrated in Fig.1(a) and (b), respectively. In the closed loop form of Fig.1(a) there is a single venue \(E\) and two market participants \(P\) and \(P’\). These participants receive market data updates via connections \(U\) and \(U’\), respectively. Responsive to information contained in those market data updates the participants execute logic, \(L\) and \(L’\), respectively, that decides whether or not to send orders to the venue \(E\). These orders are sent via connections \(O\) and \(O’\), respectively.

Regarding Fig.1(a), what happens if the path \(U-L-O\) (that of \(P\)) is shorter (i.e., takes less time to transmit and/or process information) than path \(U’-L’-O’\) (that of \(P’\))? The answer is that participant \(P\) will be advantaged in his/her trading over \(P’\) on venue \(E\) because responsive to the same information emanating from \(E\) the orders sent by \(P\) will reach the venue earlier than those of \(P’\). In a similar line of argument to that of the earlier insider trading example, \(P\) will then likely be able to buy (or sell) the resource (e.g., the stock) on \(E\) at a more favorable price than that attainable by \(P’\).

This closed-loop form of temporal fairness is instructive for two reasons. The first is that it seems to be the sum of

---

\(^1\)There are some exceptions to this that exist or are being proposed in law. E.g., in Korea where the regulator is considering outlawing microwave connections between geographically separate venues [7], and in NY State where “robo-scalping” i.e., using computers to buy concert tickets online ahead of human buyers is illegal [16].
B. Why Temporal Fairness can be Elusive

With the rise of high frequency trading, the number of messages (orders) sent into, and consequently, the number of market data updates emanating from ETVs has grown massively in recent years [21]. Consequently, to achieve adequate performance, operators of ETVs have tended to architect them so as to increasingly parallelize the operations those venues perform [14]. In many cases, the resulting architecture from this parallelization is of the form shown in Fig.2, which also roughly corresponds to that of Thomson Reuters Matching (TRM).

What the architecture in Fig.2 shows is that there are a plurality of order gateways, matching engines, and market data distributors. Roughly speaking, each order gateway accepts orders from some subset of market participants and upon inspecting the payload of each order message, determines the appropriate matching engine to which to route it. Instruments traded on the venue are divided among matching engine instances, such that each instrument belongs to exactly one such instance. Each market data distributor sends to some subset of participants on the venue their market data updates. This style of market data distribution where participants are divided among distributor instances is probably more prevalent on ETVs where instruments are traded on the basis of bilateral credit, e.g., in the spot foreign exchange (FX), and generally different, so-called credit-screened data is sent per participant even at the same market data update [22]. In other asset classes such as equities or futures that tend to be centrally cleared, the same data is sent to all participants at each market data update and fewer distribution components each servicing all participants may be employed, along with IP Multicast.

In the style of architecture shown in Fig.2 temporal fairness can be elusive for two reasons. First, consider two distinct market participants connected to two distinct order gateways. If the first participant’s gateway is ‘less loaded’ than the second participant’s, then likely the first participant’s orders will be subject to less delay at this gateway, and therefore reach the matching engine first, even if both participants have the exact same response time. If the volume of message traffic on the venue necessitates multiple gateways so the venue can achieve acceptable throughput and response times, and the load on any given gateway is not predictable at any point in time, how might the venue operator ensure temporal fairness among market participants with respect to the receipt of orders? Put another way, and with reference to Fig.1(a) (or equivalently Fig.1(b)), how given this architecture can the venue operator ensure that the times taken to transmit information over $U$ and $U'$ ‘equal to’ one another?

Second, if there are a plurality of market data distribution components each serving a subset of the market participants on a venue how—similar to the issue of varying delays in the order gateways—does one ensure that one market data distribution component is not slower than another? One solution in a bilateral credit market might be to use the hybrid IP Multicast scheme as described in [23], but even Multicast is not guaranteed to solve the problem. The Multicast discipline implemented by many commercial network switches seeks to avoid head-of-line blocking [24] [25]. The result of this is that the same Multicast packet may be offered to different output ports on the switch at different times.

The line of argument so far has been that it is difficult to ensure ‘same time’ transmission of data to and from an ETV. All of the above aside, is it even possible? The answer is probably “no” for two reasons. One is that modern computer networks serialize data. Wherever in the network’s topology there is demultiplexing to a single physical connection a total ordering of packets on that connection is induced, which implies those packets will be separated in time by at least the ‘insertion delay’ of the network. Second, instruments that trade on ETVs are not infinitely divisible. If, say, three orders could be received by the venue at the exact same time but there was only one unit of the resource available, one among the three orders would have to be chosen for matching to the resource. What could quite reasonably be construed from this is whichever one was chosen was subject to less delay than the two that were not. In other words indivisibility of a resource may also force partial or total (temporal) orderings.

C. A Stochastic Definition of Temporal Fairness

If we accept that, despite what is stated in securities laws, for any given event it is impossible to ensure same exact time dissemination of information to all market participants how might we reconcile this with the intent of those laws in the context of an ETV? The answer would seem to be by ensuring the allocation of scarce resources among participants on the ETV is the same over a long time horizon as if the same exact time constraint were feasible.

Recalling from the open-loop example of Fig.1(b) that both the transmission of data to the venue, and receipt of data from the venue are individually important (and not just important in their sum, as was illustrated in the closed-loop example of Fig.1(a)) a stochastic definition of temporal fairness for an ETV is proposed as follows. An ETV may be considered temporally fair if over some time horizon involving a plurality of distinct disclosure events and for all pairs of participants $\{P, P'\}$ and each instrument traded on it the following two conditions are met:

Fig. 2. Typical Architecture of an Electronic Trading Venue.
C1 The number of times a participant $P$ is sent market data updates before $P'$ is approximately equal to the number of times $P'$ is sent them before $P$. 

C2 When $P$ and $P'$ are competing for a certain type of resource by way of sending orders, the number of times the orders of $P$ overtake those of $P'$ is approximately equal to the number of times those of $P'$ overtake those of $P$.

With respect to C2 above, an order overtake another if it was sent earlier, but was ultimately processed by the venue (e.g., against the instrument’s limit order book, where the allocation of the resource to the order takes place) later. More concretely, and to refer back to an example from Fig.2 this may mean the first order was sent to a slower gateway and the second to a faster gateway.\(^3\)

Also with respect to the C2 above, the phrase certain type of resource is used to indicate that there are many types of resource in contention, each with different benefits for market participants associated with them. The benefit a market participant receives by having an order overtake in one type of resource may not offset by the benefit s/he forgoes by being overtaken in another distinct type of resource. In this context, even for a single instrument on a venue, a plurality of distinct resources exist e.g., buying the instrument by lifting the best offer may be one type of resource, which is distinct from achieving a position near the front of the queue at the top of the limit order book which may be another type of resource (for a discussion of the benefits of being earlier in the queue at a price level in a limit order book, see [27]).

III. IMPROVING TEMPORAL FAIRNESS

When confronted with an ETV that may not be meeting conditions C1 and C2 described above, what actions might the venue operator take towards improving its temporal fairness? One approach is to completely rearchitect it so there is one order gateway per matching engine and so that, by way of say Field Programmable Gate Array (FPGA) technology, a single gateway has sufficient performance characteristics such that can handle all participants on the venue and guarantee ordering of packets ‘on the wire’ is the same as their ordering ‘in the application’, thereby eliminating order ‘overtaking’ [28].

Another such approach may involve attempting to eliminate the transmission of data over the network by ‘moving’ the logic that otherwise existed in each market participants’ trading system into the venue itself [29], [30].

A less invasive approach—and one that is less operationally risky to implement—is to simply change the way the ETV’s matching algorithm works from being FIFO to something else that de-emphasizes the effect of being the ‘fastest’ (the cause of which per Fig.1 may be one or both of: temporal unfairness in the ETV or the social welfare aspect of participants response times). Such a mechanism could easily be retrofitted to an existing ETV by e.g., changing a small portion of the matching engine’s source code. The question then is—given that several such mechanisms for de-emphasizing the effect of speed have been proposed in the literature (see e.g., [17], [18], [33], [34])—what criteria should be used to decide which of them to use? Some criteria their respective justifications as such are provided below.

A. SOME CRITERIA FOR EVALUATING MECHANISMS

ETVs, besides being distributed systems, are also complex systems because they contain feedback loops [35]. Specifically, when properties of the venue are changed by its operator, market participants adjust their behavior to exploit those changes in a way that most benefits them. A phrase that is sometimes used, relating to how participants might exploit a change, especially in a way that circumvents its intended purpose, is gaming. A first criteria then is consideration of how a market participant might game a mechanism.

As mentioned earlier ETVs do not exist in isolation. Often-times the same instrument trades on and is fungible across a plurality of such venues. ETVs are in competition for share of the total trades that occur in the market for an instrument. When faced with decisions about where to route their orders, among other factors, market participants consider delays imposed by each ETV on their orders. Longer delays mean more uncertainty for them about the disposition of their order and therefore venues exhibiting long delays may have reduced routing priority relative to other such venues [20]. A second criteria then is length of delay the mechanism requires to achieve the desired effect.

Also relating to delays, market participants may prefer fixed delays that are known in advance over variable ones (see e.g., Rubinow of NYSE Euronext as quoted in [36], and Cifu of Virtu as quoted in [37]). A third criteria then is the variability of the delay the mechanism requires to achieve the desired effect.

Relating to the use of the phrase desired effect above, that which we desire is for two (or more) participants to all have a substantially equal probability of having their orders processed by the ETV first whenever they are competing for a common resource, to the extent those orders are received at different times due to temporal unfairness on the venue. To give a concrete example of this, if four participants are all racing to buy at (i.e., take) the prevailing offer on an instrument. When faced with decisions about where to route their orders, among other factors, market participants consider delays imposed by each ETV on their orders. Longer delays mean more uncertainty for them about the disposition of their order and therefore venues exhibiting long delays may have reduced routing priority relative to other such venues [20]. A second criteria then is length of delay the mechanism requires to achieve the desired effect.

Finally, consideration should be given to what potentially negative effects a mechanism may have on legitimate trading behaviors. For instance, would a market participant submitting a large number of orders ‘all at once’ as a result of their

\(^3\)The astute reader may have noticed Fig.2 can also be viewed as a queuing system—orders queue at gateways before being sent to matching engines. In this vein, this is reminiscent of the work of Gordon, where she notes in some situations the penalty of a “slip” (being overtaken) might be statistically offset by the benefit of a “skip” (overtaking) [26].

\(^4\)Regulators take a very dim view of ETV outages, even if those outages are inadvertent [31], [32]. It follows that an implicit criteria in selecting an approach to improve temporal fairness is its operational risk.
market-making activities be disadvantaged by a mechanism relative to other participants sending fewer orders and acting say only as takers of liquidity? A fourth criteria then is unintended consequences of the mechanism that may arise from inherent differences in legitimate trading behaviors among participants.

B. Ideal Latency Floor Mechanism

The mechanism ultimately selected for deployment on TRM in mid-2016 was the Ideal Latency Floor [33]. In short, it works by detecting the first order in a ‘race’ for a particular type of resource, and responsive to the receipt of that first order it creates a batch and timer for that race. As additional orders also belonging to that same race are received they are added to its batch until 3ms have elapsed since the receipt of the first order, at which time the orders are grouped by participant, the list of participants is randomly shuffled, and the orders are drained by participant for ultimate processing against the instrument’s limit order book.

A unique property of the mechanism is that even for a single instrument on the ETV a number of races and thus batches can be active at once. There can be a race to sell (i.e., take) the bid, a separate race to buy (i.e., take) the offer, and separate races to bid or offer (i.e., make) at each price-level in the limit order book. Each of these races relates to a distinct resource that can be in contention among a plurality of market participants: queue position at a certain price-level in the limit order book, or immediately buying (or selling) the instrument at its prevailing offer (or bid).

To see how this unique property relates to the criteria of unintended consequences consider a single market participant sending two different orders ‘at once’—one to ‘make’ by bidding higher and another one to ‘take’ by buying at the prevailing offer. If two other participants both send single orders to compete against this first participant (one bidding higher and the other taking the offer) the resource allocation will be such the first participant has an equal chance of ‘winning’ the race for each resource as the other participants. If however, there was only one batch per instrument (and not a batch per resource, as per the mechanism of [34]), then due to the grouping of orders by participant, the first participant would have a 50% chance of winning the race for one of the resources, and a 0% chance of winning the race for the other. This is also contrary to what was stated earlier as the ‘desired effect’ where all competing participants have an equal chance of winning the race for each resource in contention.

Relating to the above it might seem that, at least in part, the desired effect of the mechanism in [34] is not being met because of the grouping of orders by participant. By grouping orders in this way, participants who send more orders ‘at once’ might be perceived as being penalized relative to those who send fewer. Grouping per participant helps avoid the negative criteria of gaming though. In the mechanism of [18] speed is de-emphasized by imposing a random delay nominally between 0 and 10ms on each order before it is ultimately processed by the ETV. A participant could game this by sending multiple ‘copies’ of the same order, which would statistically increase their chances of one of those orders being subject to a 0ms delay, and therefore likely being processed first [17]. This specific form of gaming has previously been observed on a major ETV [28].

In terms of the variability of the delay criteria, the vast majority of orders sent to TRM are subject to the full 3ms delay because it is only when multiple participants actually compete for a resource that (the slower) participants orders are subject to shorter delays (recall that it is the first order in a batch/race starts its timer). In the mechanisms of [18] and [34] the delays are highly variable. In the mechanism of [17], the trading day is divided up into contiguous, fixed intervals nominally of 100ms in length, so the delay imposed depends on when relative to the end of the interval the order was received. Perhaps advantageously, by virtue of the trading day being divided up into fixed length intervals, participants will know ahead of time precisely when their order will be processed.

In terms of the length of the delay, it can be trivially shown for the chosen mechanism that length of the delay is equal to the magnitude of the temporal unfairness it seeks to nullify. If two competing participants have the exact same response times, but transmission of the data to and from the venue differs by $T$ ms, their orders will be received $T$ ms apart. If those orders are competing for the same resource, they will be put in the same batch, and hence will have the same probability of being processed first (the so-called ‘desired effect’). Other mechanisms tend to require very long delays to even approach let alone meet the desired effect. In the mechanism of [18] if a 0-10ms delay is imposed, and two orders are received say 3ms apart, then the chances of the slower order being processed by the ETV first is $24.5\% = \frac{10−3}{10} . \frac{10−3}{10} . \frac{1}{2}$. If the delay imposed is instead 0-100ms then the probability is $47.045\% (= \frac{100−3}{100} . \frac{100−3}{100} . \frac{1}{2})$, which still falls short of our desired 50%. In short, to approach the desired effect relatively long delays are required. The math is similar and conclusion is the same for the mechanism of [17], if one assumes the events triggering races are randomly distributed with respect to time.

IV. Conclusions and Future Work

This paper has shed light on the meaning of temporal fairness on ETVs, has provided reasons why widely-held views of it might be impractical, and has identified why the social welfare aspect of participants’ response times are a source of confusion. The mechanism deployed on a major spot FX venue, TRM, and criteria used in its selection have been described. Since its deployment the mechanism has been

\[ \text{Only } \sim 10\% \text{ of price-taking on TRM is contentious in that it involves two or more participants racing together (within a 3ms window); the percentage contention in price-making is even lower. } \]
well-received by participants on TRM and no unintended consequences or gaming associated with it have been observed. Some directions for future work are to formally show equivalence of resource allocation in the stochastic definition of temporal fairness with that of the same exact time form. Perhaps more fundamentally too, is to address whether the proposed stochastic definition is sufficient given its implicit assumption that a market participant’s response time is fixed and constant when in reality it is more likely to follow a distribution. This may mean that the magnitude of the delays that occur when sending data between pairs of participants—and not just ‘equalizing’ the number of times they respectively receive it first (or overtake one another)—also matters.

Finally, as was rightly noted by its anonymous reviewers this is a practice-oriented paper, and as such a future work that is longer, and more formal in its exposition especially in light of the more general works on both randomization and fairness lengthier, and more formal in its exposition especially in light.

of the more general works on both randomization and fairness

and not just ‘equalizing’ the number of times they respectively receive it first (or overtake one another)—also matters.

REFERENCES