The market nanostructure origin of asset price time reversal asymmetry

Marcus Cordi, Damien Challet, Serge Kassibrakis

The market nanostructure origin of asset price time reversal asymmetry

Marcus Cordi[1], Damien Challet[1], and Serge Kassibrakis[2]

[1]Laboratoire de Mathématiques et Informatique pour les Systèmes Complexes, CentraleSupélec, Université Paris Saclay, 3 rue Joliot-Curie, 91192, Gif-sur-Yvette, France
[2]Swissquote Bank SA, chemin de la Crétaux 33, 1196 Gland, Switzerland

December 29, 2018

Abstract

We introduce a framework to infer lead-lag networks between the states of elements of complex systems, determined at different timescales. As such networks encode the causal structure of a system, infering lead-lag networks for many pairs of timescales provides a global picture of the mutual influence between timescales. We apply our method to two trader-resolved FX data sets and document strong and complex asymmetric influence of timescales on the structure of lead-lag networks. Expectedly, this asymmetry extends to trader activity: for institutional clients in our dataset, past activity on timescales longer than 3 hours is more correlated with future activity at shorter timescales than the opposite (Zumbach effect), while a reverse Zumbach effect is found for past timescales shorter than 3 hours; retail clients have a totally different, and much more intricate, structure of asymmetric timescale influence. The causality structures are clearly caused by markedly different behaviors of the two types of traders. Hence, market nanostructure, i.e., market dynamics at the individual trader level, provides an unprecedented insight into the causality structure of financial markets, which is much more complex than previously thought.

PACS numbers PACS numbers.

1 Introduction

The collective behaviour of investors in financial markets plays a major part in shaping the complexity of price dynamics. A major challenge in the analysis and modelling of market dynamics comes from the very large heterogeneity of market participants, particularly with respect to their activity rate and feedback speed. Most agent-based models of financial markets omit timescale heterogeneity, usually focusing on strategy heterogeneity (fundamentalists, trend-followers or noise traders) and the way they learn to use them (see Hommes (2006) for a review; see however Marsili and Piai (2002); Mosetti et al. (2006); Kroujiline et al. (2016)).

The typical time-horizon of trader activity ranges from a fraction of a second to a few months (Dacorogna et al., 1998; Zumbach, 2009). A fundamental question is thus how to characterize the causal structure of market activity across timescales. In other words, is there a hierarchical (or more complex) structure in which activity propagates? Since trader-resolved data is hard to obtain, past works focused on price dynamics and volatility propagation. Intuitively, the price dynamics should reflect in some way heterogeneous trader time horizons (see e.g. Müller et al. (1993)). Early works exploit the intuitive analogy between turbulent flows and price changes (Ghashghaie et al., 1996); simple cascade models of the price dynamics have been proposed (Lux et al., 2001). Heterogeneous trader timescales may also explain why multiscale GARCH models are generally much better than plain GARCH ones (see e.g. Lynch et al. (2003); Borland and Bouchaud (2005); Chicheportiche and Bouchaud (2014)). In particular, Müller et al. (1997) argue that since coarsely-defined volatility

predicts finely-defined volatility significantly better than the other way around, the behaviour of long-term traders should influence the behaviour of short-term traders.

The above discussion implicitly assumes that prices are time reversal asymmetric (TRA). Zumbach and Lynch (2001), and Zumbach (2009) show indeed that financial time-series are significantly asymmetric with respect to the reversal of the arrow of time. While classical models of price and volatility dynamics are not TRA, GARCH processes that incorporate price returns defined over several time scales are TRA (Zumbach and Lynch, 2001; Zumbach, 2009; Chicheportiche and Bouchaud, 2014). The same holds for Hawkes processes, which are causal processes by definition and hence ideal candidates for financial modelling (Bacry et al., 2015), although their univariate and symmetric multivariate versions are surprisingly weakly TRA (Blanc et al., 2017; Cordi et al., 2018).

While there are many ways to define the timescale of a given trader, we take a more global approach here and rely instead on the notion of groups of agents determined at various timescales (seconds, minutes, hours, etc.), and investigate how the activity of one group at a given timescale influences the activity of other groups at another timescale. This opens up the possibility of inferring multi-timescale causal networks of trader activity directly instead of relying on analogies. Note that the framework which we introduce here is generic and applies to any system in which the state of one of its elements over a given time window may be summarized by a discrete state, from a small set of possible states.

Groups of traders are determined with Statistically Validated Networks (SVNs); SVNs were introduced by Tumminello et al. (2011) and have been applied e.g. to mobile communication networks (Li et al., 2014), clusters of orthologous genes, and the relationship between actors and movies (Tumminello et al., 2011). They were then used to cluster Finnish investors (Tumminello et al., 2012) and more recently to understand their long-term ecology (Musciotto et al., 2018). The main idea is that a group of similar traders should act in a similar way. The trick is to define networks of interaction according to the degree of pairwise synchronization between the actions of traders and use community detection of the resulting network to define groups of traders. Crucially, since the actions of all the members of a group are remarkably similar, the aggregate action of the group is representative of the action of each of its members, which is very helpful to reduce the dimension of trader datasets.

SVNs rely on time coarsening at a given timescale (e.g. 1 day, the best available resolution of the dataset analyzed in Tumminello et al. (2012), or 1 hour in Challet et al. (2018)). Which timescale to choose is not obvious, all the more since traders have widely different activity rates. As we shall see below, the answer depends on the type of traders (retail or institutional) and most probably on the clientele of a broker.

Recently, Challet et al. (2018) introduced Lead-Lag SVNs (LL-SVNs) to infer lead-lag networks between the states of agents in complex systems and applied them to trader-resolved data. The persistence in LL-SVNs is large enough to make it possible to predict the sign of the order flow and the VWAP of a broker clients over the next hour. A reason why these lead-lag networks exist and persist is that investors consistently react with different speeds to common information (Boudoukh et al., 1994; Jegadeesh and Titman, 1995).

Here, we extend the LL-SVN method to lead-lag networks between states determined at two different timescales. This is needed to infer how information flows from one timescale to another and to find asymmetric reciprocal influence, as is the case in trader-resolved data. Causality with respect to these two timescales is then well defined, and the ensemble of causality relationships between many pairs of timescales provides a fine picture of how information propagates in a complex system. We also discuss how the TRA of the activity of the two types of traders in our dataset compares with that of the volatility, i.e., how to relate macroscopic price properties to nanoscopic decisions, market microstructure focusing on price formation from anonymous orders sent by traders.

2 Method

2.1 SVNs and LL-SVNs

Assume that one has N time series, e.g. the transaction history of N traders. The SVN method works as follows: one first chooses a time resolution Δt and splits the time into slices of length Δt. Here, $(t, \Delta t)$ denotes the timeslice $[t, t + \Delta t[$, and for the sake of simplicity, we shall write it as t when no ambiguity arises.

For each timeslice, one summarizes the activity of each timeseries by a discrete state taken from a small number of possible states. For traders, it is natural to define four different states: mostly buying (+1), mostly selling (-1), neutral (0) and inactive (NA). The imbalance ratio of time series i for each timeslice t is then

$$\rho_i(t) = \frac{v_i(t)}{a_i(t)} \tag{1}$$

where $v_i(t) = v_i(t, \Delta t)$ is the total signed transaction volume of trader i during timeslice t, and, similarly, $a_i(t)$ is the sum of the absolute trading volume during this timeslice t. The state of agent i during timeslice t is

$$\sigma_i(t) = \begin{cases} 1 & \text{if } \rho_i(t) > \rho_0 \\ -1 & \text{if } \rho_i(t) < -\rho_0 \\ 0 & \text{if } \rho_i(t) < |\rho_0| \\ \text{NA} & \text{if } v_i(t) = a_i(t) = 0. \end{cases}, \tag{2}$$

As in previous works, we use $\rho_0 = 0.01$, but the specific choice of this parameter does not have much influence on the results provided that it is small.

The level of synchronicity between two given states of two given traders is determined by assuming that the occurrence of each state follows a Poissonian process in discrete time (the timeslices). Then, using an exact expression for the probability of synchronicity of two independent process, it is straightforward to compute the p-value of these states for these traders. The computation is performed for all possible pairs of traders and all allowed pairs of states. Here, since one wishes to group traders, the set of allowed pairs is $(\{(1,1),(-1,-1),(0,0)\}$; we drop the inactive state by focusing on the most active traders. Testing all the pairs of traders for each possible state pair yields a large number of tests, thus multiple hypothesis testing correction is needed: we use the False Discovery Rate (FDR) (Benjamini and Hochberg, 1995), with an FDR rate set to $p_0 = 0.05$. An SVN network is obtained by keeping links whose p-values are smaller than the FDR-adjusted threshold (see Tumminello et al. (2012) for more details).

The resulting network may then be decomposed into groups (communities) by using the InfoMap method (Rosvall and Bergstrom, 2008), which is one of the most efficient methods of community detection in networks (Lancichinetti and Fortunato, 2009). The multi-links are converted into weighted links by assigning a weight equal to the number of validated links between two traders. Since links are only allowed between traders who take similar actions, the state of each group of traders is well defined and mirrors those of the traders that it includes.

Let us introduce some more mathematical notations. Mathematically, one can define the state of group $g \in G$, where G is the set of all groups, during timeslice t, by

$$\sigma_g(t) = \begin{cases} 1 & \text{if } \rho_g(t) > \rho_0 \\ -1 & \text{if } \rho_g(t) < -\rho_0 \\ 0 & \text{if } \rho_g(t) < |\rho_0| \\ \text{NA} & \text{if } V_g(t) = 0, \end{cases} \tag{3}$$

where $\rho_g(t) = \frac{V_g(t)}{|V_g(t)|}$ and $V_g(t) = \sum_{i \in g} v_i(t)$ is the aggregate signed volume of the traders belonging to group g during timeslice t. We also define $A_g(t) = \sum_{i \in g} a_i(t)$ as the aggregate absolute volume of the traders belonging to group g during timeslice t. Grouping traders is surprisingly efficient and

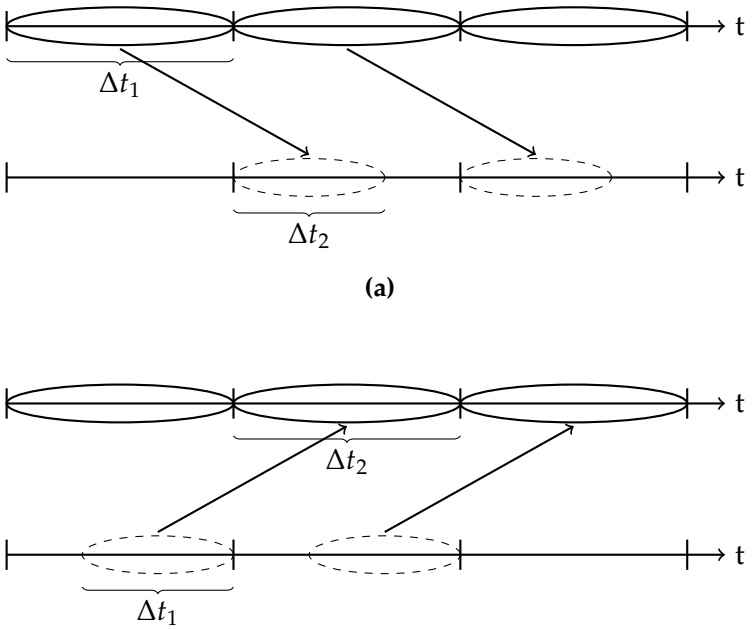

Figure 1: Schematic diagram showing how lead-lag links are established when (a) $\Delta t_1 > \Delta t_2$ and (b) $\Delta t_1 < \Delta t_2$. The dotted lines indicate that the state of the group has been recalculated if $t/\Delta t_2$ is not an integer, which corresponds to a time-shift with regards to the time-interval where the groups were determined.

significantly decreases the dimensionality of the problem, i.e., the effective number of timeseries to track in a population of clients of a broker. As it reduces the dimension of the data set, using groups tremendously helps to simplify and speed up the determination of lead-lag networks, and we shall keep this procedure.

The easiest case is of course when the states of traders who lead and who lag are determined with the same coarse resolution Δt, as in Challet et al. (2018).

2.2 LL-SVNs with two timescales

The main methodological contribution of our work is to introduce a general framework to infer lead-lag relationships between groups whose states are determined at two (possibly) different timescales.

The general principle is simple (see Fig. 1 for a graphical illustration):

1. Let Δt_1 and Δt_2 be two timeslice durations.
2. Apply the SVN method to both Δts in order to determine two sets of groups G_1 and G_2 (optional but recommended)[1].
3. Find SVNs between the suitably lagged values of group (or agent) states.

By convention, in the following, Δt_1 is the timescale at which the leading states of agents are determined and Δt_2 the timescale of the lagging states of agents. When $\Delta t_1 = \Delta t_2$, the segmentations of a timeseries for both the leading and lagging states coincide and no special caution regarding their alignment is needed. However, when $\Delta t_1 \neq \Delta t_2$, for a given time $t = k_1 \Delta t_1$, $k_1 \in \mathbb{N}$, the boundaries of timeslices for both timescales are generally not aligned, i.e., there is generally no integer k_2 such that $k_2 \Delta t_2 = k_1 \Delta t_1$. This is a problem when inferring LL-SVNs, as non-aligned slices induce a lag between the end of the leading slice and the lagging one, which would then reduce the strength of causality relationships. In addition, one needs to avoid computing the states of agents

[1]When the number of agents is not too large, this step may be skipped.

or groups on partially overlapping timeslices for the longest timescale. This is why we align the computation of the states at times $t = k \max(\Delta t_1, \Delta t_2)$, $k \in \mathbb{N}$ (see Fig. 1).

This alignment problem suggests several possible variations, both in terms of how and when groups and their states are determined. Methods I, II, and III introduced below each define a set of lead-lag links.

2.2.1 Method I

This method only uses a single grouping of agents, and is thus both faster and simpler. While agent grouping (clustering) is done with respect to one of the two timescales (see below), the state of each group is computed in timeslices of length Δt_1 and Δt_2 which are aligned as in Fig. 1. Only clustering with respect to a single timescale may sometimes miss subtle differences of group membership, especially if the timescales are very different.

Method I works as follows when $\Delta t_1 > \Delta t_2$ (it is assumed here that $t = k\Delta t_1$, where $k = 0, 1, 2, \ldots$):

1. Time is discretized at timescale Δt_1 in order to obtain the group set G.
2. For each group $g \in G$ and timeslices $(t, \Delta t_1)$, the states $\sigma_g(t, \Delta t_1) = \sigma_g^{(1)}(t)$ are determined.
3. For each group $h \in G$ and timeslices $(t, \Delta t_2)$, the states $\sigma_h(t, \Delta t_2) = \sigma_h^{(2)}(t)$ are determined.
4. For each possible pair (g, h), g and $h \in G$, the p-value of the synchronicity between $\sigma_g^{(1)}(t)$ and $\sigma_h^{(2)}(t + \Delta t_1)$ is calculated.

When $\Delta t_1 < \Delta t_2$, one needs to consider $t = k\Delta t_2$, where $k = 0, 1, 2, \ldots$:

1. Time is discretized at timescale Δt_2 in order to obtain the group set G.
2. For each group $g \in G$ and timeslices $(t, \Delta t_2)$, the states $\sigma_g(t, \Delta t_2) = \sigma_g^{(2)}(t)$ are determined.
3. For each group $h \in G$ and timeslices $(t - \Delta t_1, \Delta t_1)$, the states $\sigma_h(t - \Delta t_1, \Delta t_1) = \sigma_h^{(1)}(t - \Delta t_1)$ are determined.
4. For each possible pair (g, h), g and $h \in G$, the p-value of the synchronicity between $\sigma_g^{(2)}(t)$ and $\sigma_h^{(1)}(t - \Delta t_1)$ is calculated.

Since there is only one set of groups, defining self-referential lead-lag links (from one group to itself) is straightforward.

In the implementation of the method above it is clear that the time discretization used for the group classification is always based on the longer timescale, regardless of whether it acts as lead or lag. We have also implemented the method above with the shorter timescale as basis for time discretization used for the group classification, and we checked that the results did not differ significantly.

2.2.2 Method II

This method defines two groups, one for each time scale, over the whole calibration window, denoted by G_1 and G_2. Method II works as follows when $\Delta t_1 > \Delta t_2$, assuming that $t = k\Delta t_1$, where $k = 0, 1, 2, \ldots$:

1. Time is discretized at timescale Δt_1 and Δt_2 in order to obtain G_1 and G_2.
2. For each group $g \in G_1$ and timeslices $(t, \Delta t_1)$, the states $\sigma_g(t, \Delta t_1) = \sigma_g^{(1)}(t)$ are determined.
3. For each group $h \in G_2$ and timeslices $(t, \Delta t_2)$, the states $\sigma_h(t, \Delta t_2) = \sigma_h^{(2)}(t)$ are determined.

4. For each possible pair (g,h), $g \in G_1$ and $h \in G_2$, the p-value of the synchronicity between $\sigma_g^{(1)}(t)$ and $\sigma_h^{(2)}(t+\Delta t_1)$ is calculated.

When $\Delta t_1 < \Delta t_2$ the method works as follows, assuming that $t = k\Delta t_2$, where $k = 0, 1, 2, \ldots$:

1. Time is discretized at timescale Δt_1 and Δt_2 in order to obtain G_1 and G_2.
2. For each group $g \in G_2$ and timeslices $(t, \Delta t_2)$, the states $\sigma_g(t, \Delta t_2) = \sigma_g^{(2)}(t)$ are determined;
3. For each group $h \in G_1$ and timeslices $(t - \Delta t_1, \Delta t_1)$, the states $\sigma_h(t - \Delta t_1, \Delta t_1) = \sigma_h^{(1)}(t - \Delta t_1)$ are determined.
4. For each possible pair (g,h), $g \in G_2$ and $h \in G_1$, the p-value of the synchronicity between $\sigma_g^{(2)}(t)$ and $\sigma_h^{(1)}(t - \Delta t_1)$ is calculated.

Since the alignment follows the time slices of the longer timescale, we avoid any overlap (and thus unnecessary correlation) between two adjacent time slices.

2.2.3 Method III

Finally, we introduce Method III which ensures that group inference and group states are computed in the same time slices. More specifically, what is different in this method is that the groups for the shorter time-interval are determined 'in place' with regards to how their trade volumes are aggregated, depending on if the shorter timescale acts as lead or lag. We therefore have two different sets of groups for the shorter timescale in order to avoid overlap. The advantage of this method is thus that we avoid the re-calculation of the group states (which is necessary in the other two methods) and that clustering fully corresponds to the states used to determine the LL-SVN. The disadvantage is that clustering is performed with fewer events for the shorter timescale.

The method works as follows when $\Delta t_1 > \Delta t_2$ assuming that $t = k\Delta t_1$, where $k = 0, 1, 2, \ldots$:

1. Time is discretized at timescale Δt_1 in order to obtain G_1.
2. Time is discretized as $[t, t + \Delta t_2[$ in order to obtain G_2.
3. For each group $g \in G_1$, the states $\sigma_g(t, \Delta t_1) = \sigma_g^{(1)}(t)$ are determined.
4. For each group $h \in G_2$, the states $\sigma_h(t, \Delta t_2) = \sigma_h^{(2)}(t)$ are determined.
5. For each possible pair (g,h), $g \in G_1$ and $h \in G_2$, the p-value of the synchronicity between $\sigma_g^{(1)}(t)$ and $\sigma_h^{(2)}(t+\Delta t_1)$ is calculated.

When $\Delta t_1 < \Delta t_2$ the method works as follows assuming that $t = k\Delta t_2$, where $k = 0, 1, 2, \ldots$:

1. Time is discretized at timescale Δt_2 in order to obtain G_2.
2. Time is discretized as $[t - \Delta t_1, t[$ in order to obtain G_1.
3. For each group $g \in G_2$, the states $\sigma_g(t, \Delta t_2) = \sigma_g^{(2)}(t)$ are determined.
4. For each group $h \in G_1$, the states $\sigma_h(t - \Delta t_1, \Delta t_1) = \sigma_h^{(1)}(t - \Delta t_1)$ are determined.
5. For each possible pair (g,h), $g \in G_2$ and $h \in G_1$, the p-value of the synchronicity between $\sigma_g^{(2)}(t)$ and $\sigma_h^{(1)}(t - \Delta t_1)$ is calculated.

3 Dataset

Our datasets contain trader-resolved transactions of the EUR/USD currency pair and come from two independent sources: Swissquote Bank SA (SQ hereafter), a Swiss broker-dealer with a large market share in foreign exchange (FX) transactions in Switzerland, and a large anonymous dealer bank which serves major institutional clients. Both datasets list all the trades of their clients: traded currency pair, anonymous client identification number, trade time (at a millisecond resolution), signed volume, and the FX transaction rate. We focus on the EUR/USD pair as it is one of the most traded pairs in both datasets. A summary of the datasets structure and contents is provided in Table 1.

Dataset	Timespan	Traders	Trades
LB	01 Jan. 2013 - 15 Sep. 2014	$> 10^3$	$> 10^5$
SQ	01 Jan. 2014 - 30 Jun. 2014	$> 10^3$	$> 10^5$

Table 1: Basic statistics of the datasets studied for EUR/USD currency pair

While FX markets never close, transactions are quite rare during nights and week-ends. We thus focus on active hours, i.e., from 9:00 to 17:00 on week days. We only look for links between adjacent timeslices on the same day in order to avoid spurious boundary effects or overnight lead-lag links.

4 Results

Because the active population in both datasets evolves much faster than the total duration of the datasets, one cannot use the whole datasets to infer lead-lag networks. We use here rolling calibration time windows of $T_{\text{in}} = \{30, 60, 90, 120\}$ business days[2]. For each time window, we apply Methods I, II, and III to each pair of timescales Δt_1 and Δt_2 belonging to the arithmetic sequence from 5 minutes to 240 minutes (4 hours) with a step of 5 minutes (which corresponds to 1176 unique pairs of timescales). Computations over the whole length of a single dataset last for about a day for each T_{in} and each dataset using 72 cores, for all pairs of timescales. In order to speed-up computations, we only keep the 500 most active traders in each calibration windows.

4.1 $\Delta t_1 = \Delta t_2$

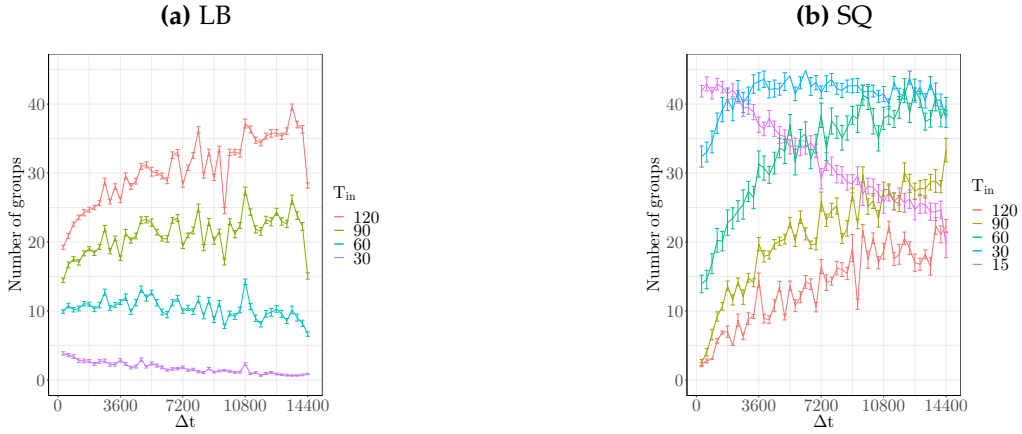

Figure 2: Average number of groups as a function of Δt and T_{in}. $\Delta t = \Delta t_1 = \Delta t_2$.

We first focus on the diagonal $\Delta t_1 = \Delta t_2 = \Delta t$. In this case, determining lead-lag networks does not require any special care and indeed the three methods defined above are identical and correspond to the single-timescale method of Challet et al. (2018). A systematic investigation of global

[2]We have also used $T_{\text{in}} = 15$ for the SQ dataset.

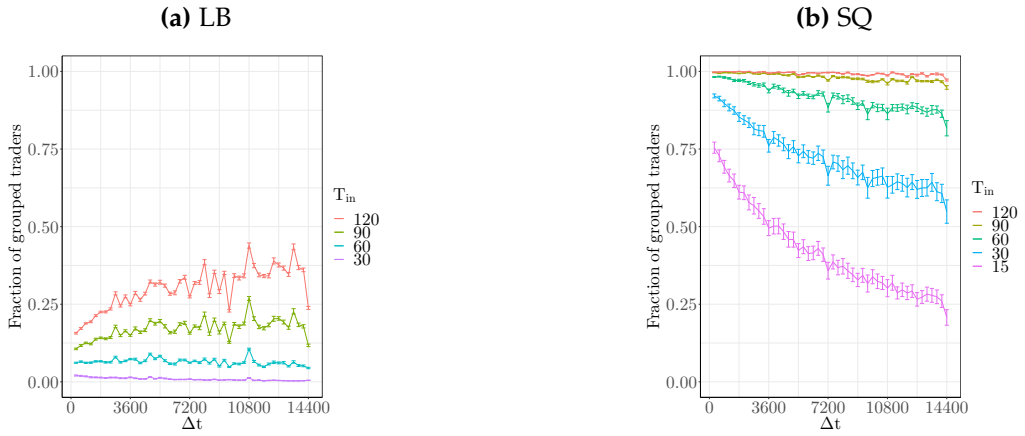

Figure 3: Average fraction of traders grouped by SVNs as a function of Δt and T_{in}. $\Delta t = \Delta t_1 = \Delta t_2$.

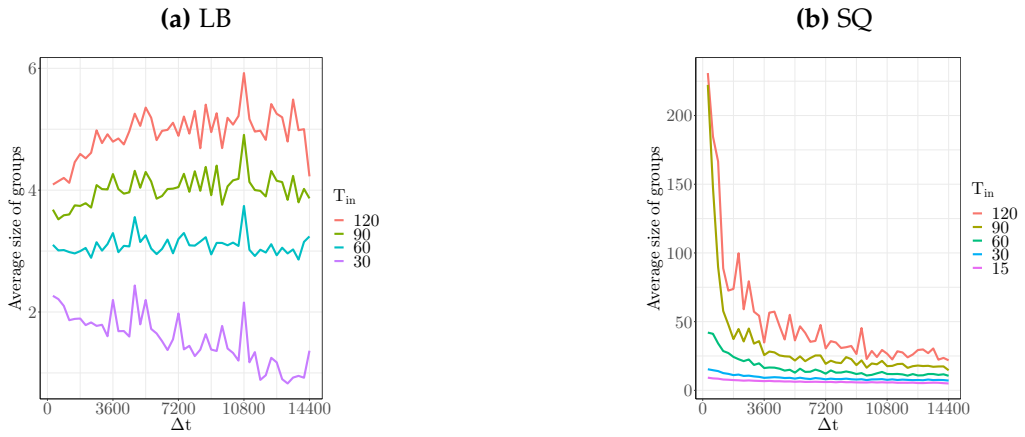

Figure 4: Average size of groups as a function of Δt and T_{in}. $\Delta t = \Delta t_1 = \Delta t_2$.

properties of lead-lag networks as a function of Δt and the window calibration length T_{in} in our datasets is necessary, as it indeed reveals fundamental differences between retail and institutional clients (at least in our datasets), which in turn will help to understand the results with two different timescales.

Fig. 2 plots the number of groups averaged over all calibration windows as a function of Δt for all T_{in}, for both LB and SQ. The number of groups found by the LL-SVNs and InfoMap is a measure of the statistically validated diversity of behaviour and of the potential richness of connectivity. For example, only a few groups of LB clients for $T_{\text{in}} = 30$ and large Δt are detected, while the largest value of $T_{\text{in}} = 120$ yields the most groups for LB. One also sees a sudden drop of the number of groups for $\Delta t = 14400s = 4h$, which is likely a by-product of the fact that we keep 8 hours of trading each day.

The number of groups of SQ retail clients behaves in the exactly opposite way unless T_{in} is small: the smaller T_{in}, the larger the number of groups. The case $T_{\text{in}} = 15$ for SQ shows that the effective number of points, proportional to $T_{\text{in}}/\Delta t$, must be large enough for the method to be powerful enough.

The group size distribution is very skewed: for example the median size of the groups is much smaller that the average group size. In fact, one often sees the emergence of a very large large group for small Δt, while other groups are typically very small. We will thus focus on $T_{\text{in}} = 120$ for LB and $T_{\text{in}} = 30$ for SQ.

The *raison d'être* of calibration in sliding windows is *a priori* the non-stationarity not only of the population of traders, but also of their behaviour. If both are roughly stationary, a longer T_{in}, at fixed Δt, should give more precise and richer results, and inversely. This is likely a major cause of the difference between SQ and LB traders, the latter behaving in a much more stationary manner.

Let us now turn to the links themselves. Since we deal with lead-lag networks, they are directed.

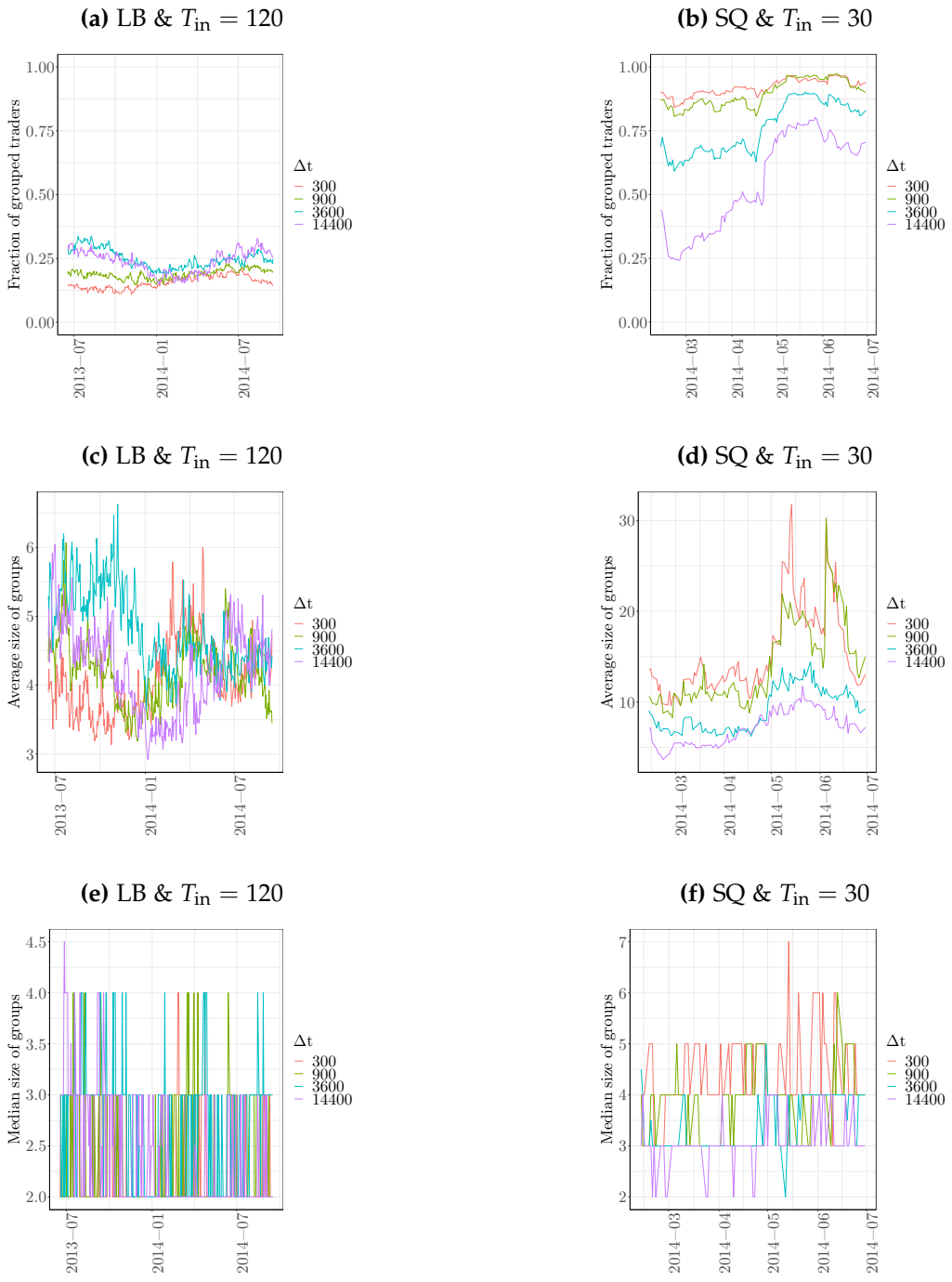

Figure 5: Fraction of traders grouped by SVNs, average size of groups and median size of groups as a function of time for LB and SQ. $\Delta t = \Delta t_1 = \Delta t_2$.

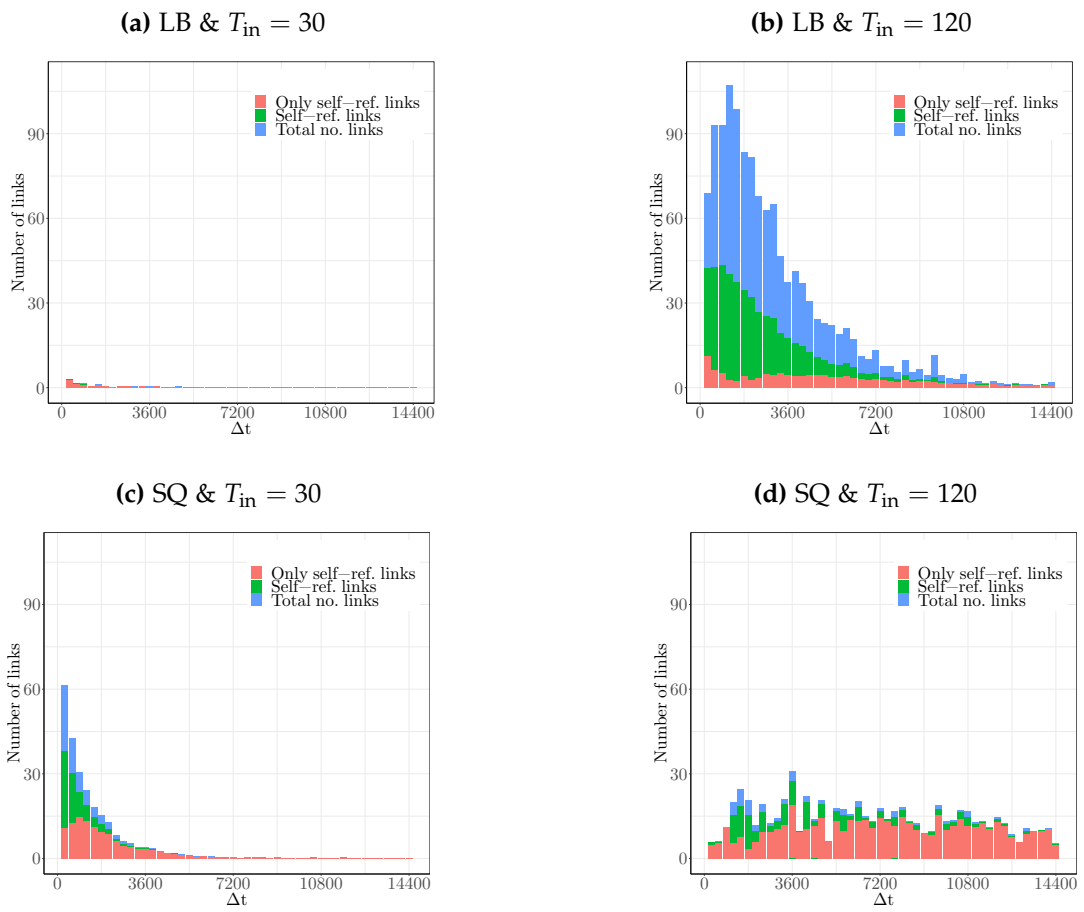

Figure 6: Total number of links as a function of the timeslice duration (in seconds) $\Delta t = \Delta t_1 = \Delta t_2$, for groups with only self-referential links, self-referential links and links to other groups and only links to other groups.

Links can be of two types: either from one group to another one, or to the same group, which we call a self-referential link. Occasionally, some groups only link to themselves, which would happen if they use an effective strategy whose activity does not systematically lead another one, but whose activity, on average, occurs at a scale comparable to Δt.

Fig. 6 plots the average total number of lead-lag links, and distinguishes within these links the average number of self-referential and 'only' self-referential lead-lag links. The lead-lag networks of the two types of traders are clearly different: the typical fraction of groups with only self links is small for LB traders, but much larger for SQ traders. The timeslice length Δt influences the number of non-self-referential links for both populations: their number decreases sharply when $\Delta t > 1$ hour and are negligible at resolutions coarser than 2 hours for SQ and 3 hours for LB.

4.2 $\Delta t_1 \neq \Delta t_2$

4.2.1 Links

When $\Delta t_1 \neq \Delta t_2$, both timescales may influence each other in an asymmetric way. Our strategy is to capture such an asymmetry by using several quantities related to both the directed network structure and the rate of trading. Each quantity is estimated for each pair $(\Delta t_1, \Delta t_2)$, each of them ranging from 5 minutes to 4 hours (14440 seconds) by steps of 5 minutes, which gives 1176 unique pairs. Since we measure these quantities over many calibration windows, we obtain a timeseries for each quantity and for each pair.

Let us first start with the number of links. The left hand side plots of Figs. 7 and 8 show the average number of links for each pair of timescales. Let us clarify the convention: Δt_1 (on the x-axis) leads on Δt_2 (on the y-axis): as a consequence, points above the $y = x$ line correspond to smaller timescales leading on longer timescales, and inversely. It is useful to keep in mind that on the diagonal $\Delta t_1 = \Delta t_2$ (Fig. 6) the number of links is maximal for small values of Δt for both LB and SQ.

The three methods give qualitatively similar results, although it is more difficult for Method III to detect links for timescales very far from the diagonal for LB. In accordance with Fig. 6, there are more links for smaller values of Δt_1 and Δt_2 around the diagonal. In addition, one generally finds that the number of links has a local maximum on the diagonal. There are also more links for some particular values of either Δt_1 or Δt_2, e.g. multiples of full hours. This may indicate that some traders have a typical activity change over 1 hour, e.g. a trading strategy that depends on the time of the day, or that they trade between, say, 9:00 and 10:00, 10:00 and 11:00, and so on.

At least for LB, it is obvious that there are more links above than below the diagonal, which implies that there are on average more links from shorter timescales to longer timescales. The statistical significance of this difference is assessed in the following way: let us denote the number of links of the pair $(\Delta t_1, \Delta t_2)$, the first timescale of the pair leading on the second one, in calibration window i by $W_i(\Delta t_1, \Delta t_2)$. One then applies a t-statistics to the timeseries of the difference $\delta W_i(\Delta t_1, \Delta t_2) = W_i(\Delta t_1, \Delta t_2) - W_i(\Delta t_2, \Delta t_1)$, In order to avoid too many false positives, we use a false discovery rate (FDR) correction for multiple hypotheses made in this plot, setting the rate at 0.2. Right columns of Figs 7 and 8 plots the selected t-stats of $\delta W_i(\Delta t_1, \Delta t_2$: blue zones correspond to lead-lag links from shorter to longer timescales, and reversely for red zones.

The plots for LB are overwhelmingly blue: there are more links from short timescales to long timescales, for the three methods. There is a clear exception for $\Delta t_1 = 4h$, which once again is probably a by-product keeping exactly 8 hours of data each day. One notes however small red regions when two groups of traders are used (Methods II and III): at around (3h, 2h) for Method II and III, for relatively small values of the lagging timescales for Method III, and (1h, 5m) and (2h, 5m) for Method II.

For the SQ traders, the link structure is much more complex. Focusing on the common results between the three methods, one finds a zone where longer timescales have more links to shorter timescales when $\Delta t_1 <$ 1h, and also around (3h, 1.5h). One also notes an alternance of positive and negative vertical stripes.

The number of links themselves are not sufficient to characterize the lead-lag between timescales

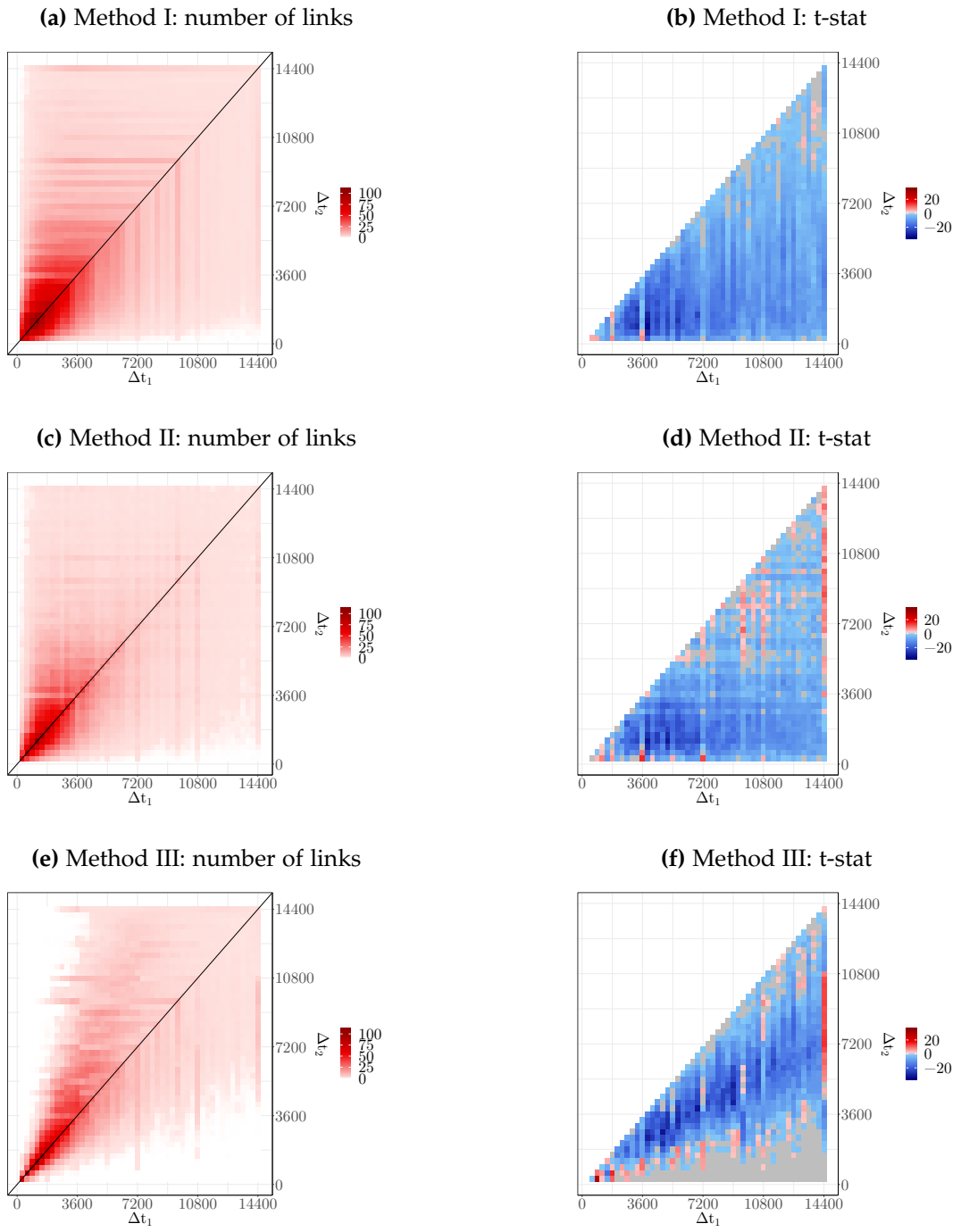

Figure 7: Left hand-side plots: average number of lead-lag links for LB (Δt_1 leads on Δt_2). Right hand-side plots: t-statistics of the difference between the number of links of the pairs $(\Delta t_1, \Delta t_2)$ and $(\Delta t_2, \Delta t_1)$; negative values indicate that shorter timescales link significantly more to longer timescales. $T_{\text{in}} = 120$

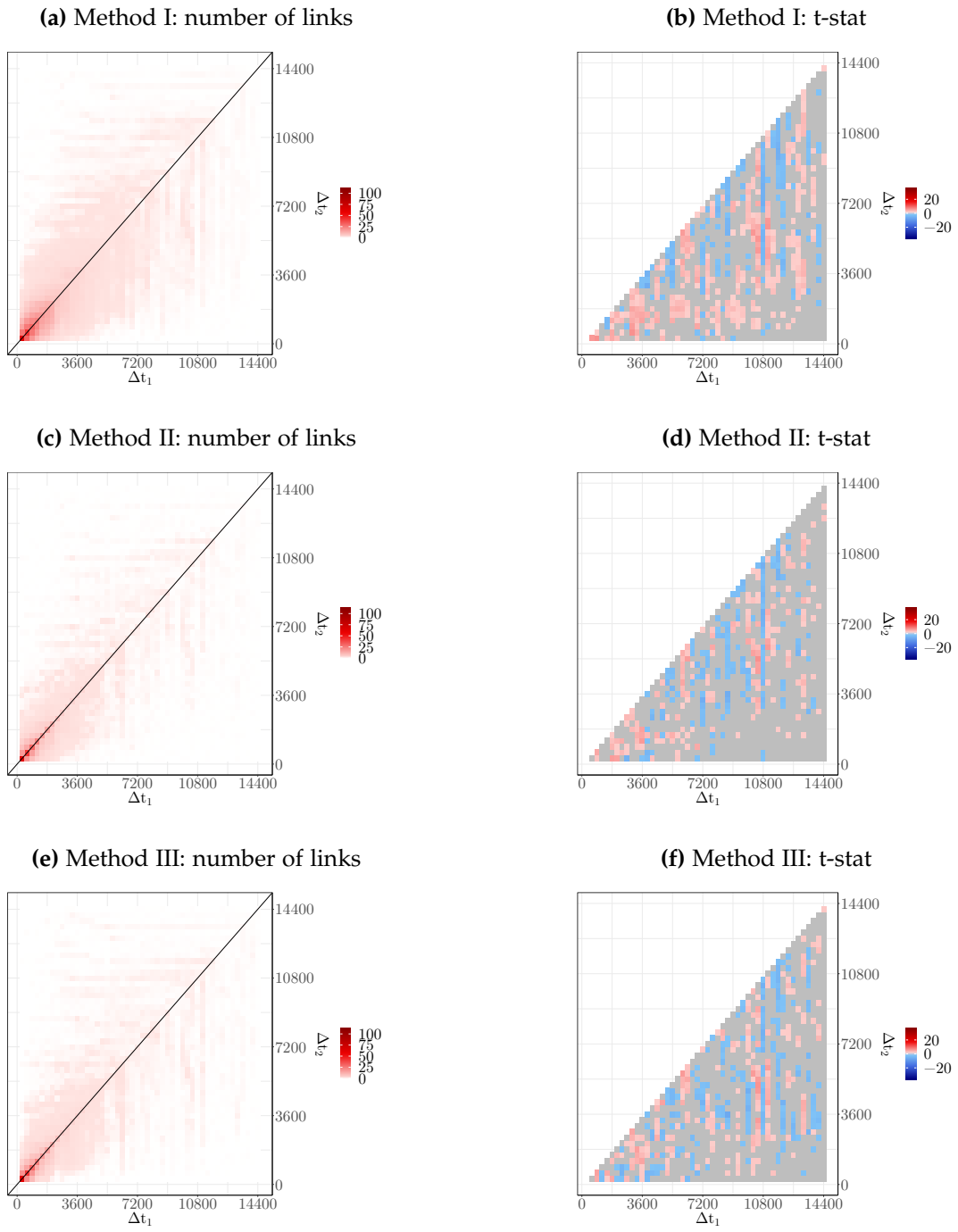

Figure 8: Left hand-side plots: average number of lead-lag links for SQ (Δt_1 leads on Δt_2). Right hand-side plots: t-statistics of the difference between the number of links of the pairs $(\Delta t_1, \Delta t_2)$ and $(\Delta t_2, \Delta t_1)$; negative values indicate that shorter timescales link significantly more to longer timescales. $T_{\text{in}} = 30$.

for traders. For example, how a given group links to other ones may also be surprising. Indeed, it is quite possible that a group has more than one link to another group, even for the same initial state. For example, group 1 may have links $+1 \to +1$ and $+1 \to -1$ with group 2. This happens quite often but is not as strange as it may appear at first: such dual links means in the case that the mostly buying activity of group 1 triggers either $+1$ or -1 in group 2. In other words, it triggers a directional activity of group 2, whose sign is undeterminate. In a prediction setting, dual links of course reduce the prediction power, but as long as enough single links do exist, order flow prediction is possible, as shown by Challet et al. (2018).

4.2.2 Activity

Being able to account for two timescales makes it possible to connect Time Reversal Asymmetry (TRA) at the level of trader behaviour to that of the price itself. TRA of prices, while being totally intuitive in financial markets, is not totally trivial to measure owing to the amount of noise in financial data. Zumbach and Lynch (2001) proposed to measure the asymmetry between historical volatility measured over Δt_h in the past and realized volatily, estimated over Δt_r. More precisely, for a given t, one estimates the historical volatility $v_h(t)$ in the interval $]t - \Delta t_h, t]$ and the realized volatility $v_r(t)$ in the interval $[t, t + \Delta t_r[$; then one estimates the correlation of v_h and v_r for all chosen ts, denoted by $\rho(\Delta t_h, \Delta t_r)$. This results in volatility correlation mugshots in which one clearly sees the asymmetry of $\rho_v(\Delta t_h, \Delta t_r)$ with respect to the diagonal $\Delta t_r = \Delta t_h$. Zumbach (2009) investigates further the TRA of volatility and proposes two more measures of TRA by noticing that the price returns in the time intervals over which volatility is estimated can be defined according to their own timescale, whose fine structure is investigated, e.g. in Chicheportiche and Bouchaud (2014).

Connecting agent activity and volatility is natural if time subordination holds (Clark, 1973). In other words, if the volatility per trade is locally constant, then the volatility in a time interval depends on the number of trades occurring in that period of time assuming that prices are diffusive. While this neglects jumps of various origins, e.g. microstructural noise due to heavy-tailed distribution gaps in limit order books (Gillemot et al., 2006), we only need to assume that there is a monotonic average relation between the number of trades and volatility to connect trader activity and volatility.

Therefore, we can estimate the correlation between the activity rate of traders in leading groups and lagging groups, determined at two different timescales, as above. Let us therefore denote the total number of trades of agents in group g during timeslice $(t, \Delta t)$ by $N^{(g)}(t, \Delta t)$; in addition, we denote by

$$N_1(t) = \sum_{g \in G_1} N^{(g)}(t, -\Delta t_1)$$

the total activity of the leading groups at time t, and similarly

$$N_2(t) = \sum_{g' \in G_2} N^{(g')}(t, \Delta t_2)$$

the total activity of the agents in the lagging groups (note that with Method I, $G_1 = G_2$). We then can compute the correlation between activity rates $N_1(t)/\Delta t_1$ and $N_2(t)/\Delta t_2$, denoted by $\rho(\Delta t_1, \Delta t_2)$.

Figs. 9 and 10 plot $\rho(\Delta t_1, \Delta t_2$ for the LB and SQ datasets respectively (left-hand side plots), and correspond to the mugshots of Zumbach and Lynch (2001) but for activity rates. In the case of LB, the asymmetry is clear and is confirmed by the right-hand side plots which report the t-statistics of $\delta\rho(\Delta t_1, \Delta t_2) = \rho(\Delta t_1, \Delta t_2) - \rho(\Delta t_2, \Delta t_1)$; only the values validated by FDR are in color, the unvalidated ones being reported in gray. As the three methods point to the same conclusion: activity on shorter timescales in the past is more correlated with future activity on longer timescales than the opposite (blue zone), this globally mirrors the dependence between the number of links and the correlation. Note that this is an anti-Zumbach effet. Methods I and II however suggest a subtler picture: the Zumbach effect emerges (red zone) when $\Delta t_2 > 2$hours. Our dataset is not sufficiently long or dense to report results for Δt_1 or $\Delta t_2 > 4$hours. The SQ dataset only leads to statistically validated TRA in a few vertical stripes, which suggests that SQ traders react to volatility

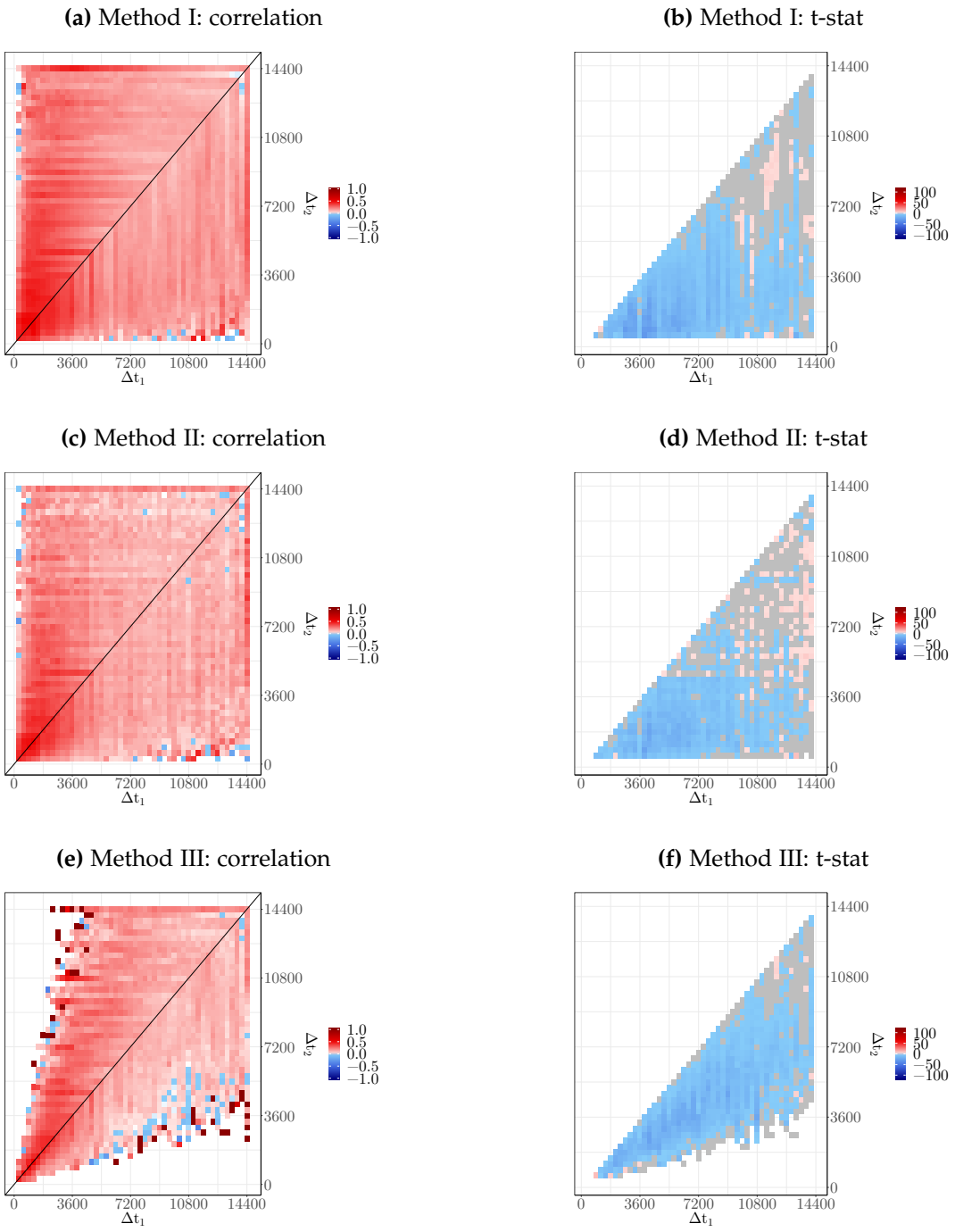

Figure 9: Left hand-side plots: average correlation between the leading (at timescale Δt_1) and lagging (at timescale Δt_2 activity rates, $\rho(\Delta t_1, \Delta t_2)$. Right hand-side plots: t-statistics of the difference $\rho(\Delta t_1, \Delta t_2) - \rho(\Delta t_2, \Delta t_1)$; negative value correspond to activity at small timescales being more correlated to future activity at larger timescale than reversely. LB dataset.

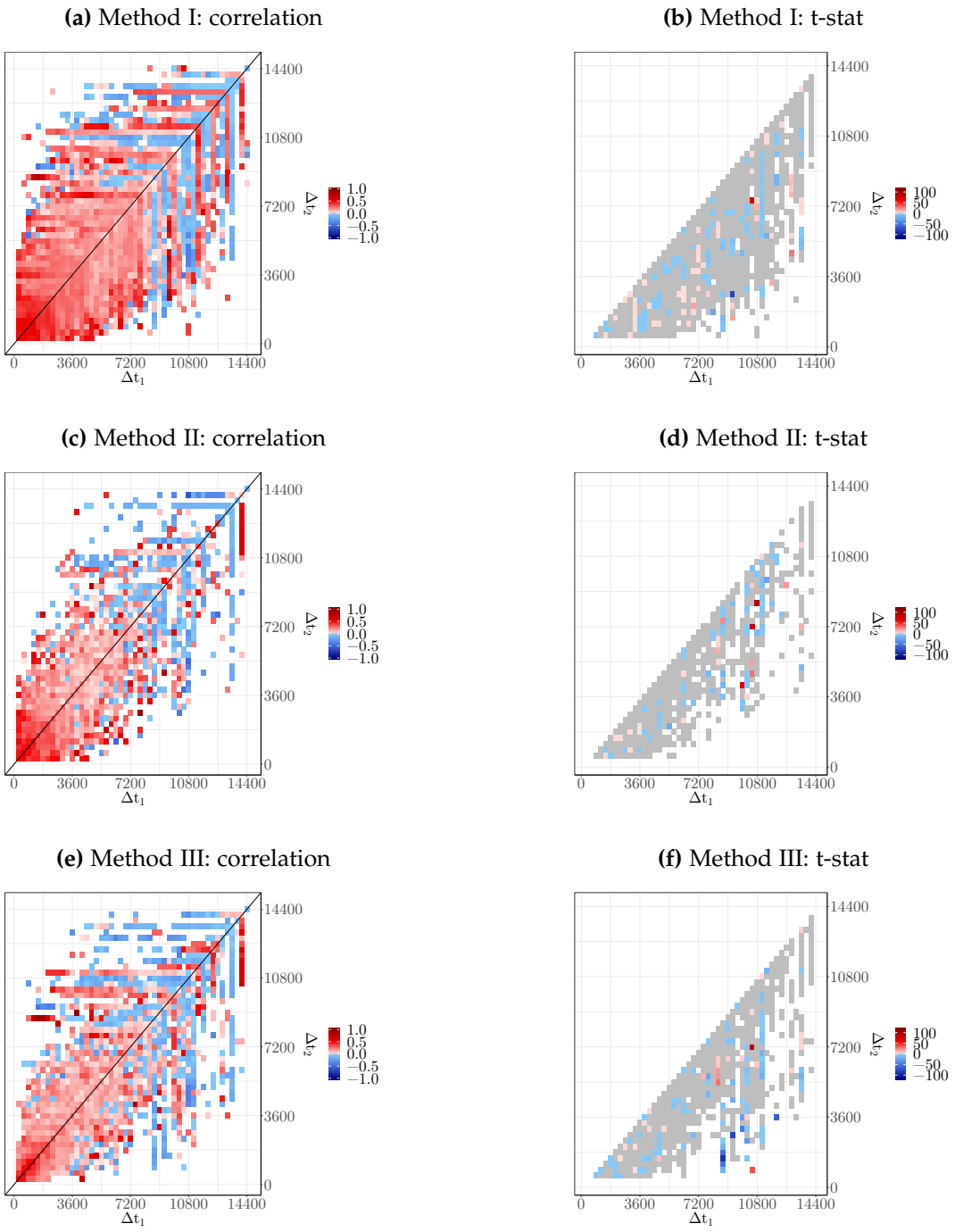

Figure 10: Left hand-side plots: average correlation between the leading (at timescale Δt_1) and lagging (at timescale Δt_2) activity rates, $\rho(\Delta t_1, \Delta t_2)$. Right hand-side plots: t-statistics of the difference $\rho(\Delta t_1, \Delta t_2) - \rho(\Delta t_2, \Delta t_1)$; negative values correspond to activity at small timescales being more correlated to future activity at larger timescale than reversely. SQ dataset.

measured at specific timescales. Comparing our results to those of Zumbach and Lynch is not straightforward because because time units are not the same: we use physical time and Zumbach and Lynch work in business time in which the activity rate is roughly constant. Hence, we also computed the mugshot of volatility computed from 5-seconds returns both for the LB dataset and the mid price infered from tick data dukaskopy.com (DK henceforth) for the same period. It turns out that the asymmetry of price volatility in LB data has the same sign as the that of activity rate between statistically validated groups, i.e., also leads to an anti-Zumbach effect. However, when using DK data, we found a Zumbach effect for $\Delta_1 < 1$ hour even in physical time. Since the traders in LB dataset are not active in all the slices of 5 seconds, it means that the LB trader activity pattern is not random but is restricted to time slices in a such way that an anti-Zumbach effect emerges.

5 Conclusions

Lead-lag SVNs between two timescales yields a fine-grained picture of the causal structure of activity in complex systems. It also opens up the possibility of understanding the time reversal asymmetry at a global level from causality between agents.

When applying this method to trader-resolved data, we found markedly different behaviors between institutional and retail traders (at least in our datasets). For example, the calibration window length at which our method detects most groups and links is much smaller for retail clients and the lead-lag network structure of the latter is more self-referential.

Despite the fact that volatility TRA in business time and physical time for small enough timescales clearly points to a larger influence from larger past timescales to smaller future timescales, the trader activity TRA and volatility TRA, when conditioned on the times at which the traders are active, are much more complex: the behavior of each category of trader corresponds to a quite specific causality structure which is neither as simple as that of volatility, nor always of the same sign. This also suggests a more intricate structure of interaction between timescales and between the types of traders than that of volatility. The latter, being the outcome of aggregation between many brokers, hence, between many types of traders, hides much of the variety of trader behavior. In passing, not having a complete trader-resolved dataset made it possible to bring to light this kind of heterogeneity. Finally, our results emphasize the richness of the dynamics of financial markets, the fundamental importance of timescales, and provides a new set of stylized facts against which agent-based models should be tested.

References

E. Bacry, I. Mastromatteo, and J.-F. Muzy. Hawkes processes in finance. *Market Microstructure and Liquidity*, 1(01):1550005, 2015.

Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the royal statistical society. Series B (Methodological)*, pages 289–300, 1995.

P. Blanc, J. Donier, and J.-P. Bouchaud. Quadratic Hawkes processes for financial prices. *Quantitative Finance*, 17(2):171–188, 2017.

L. Borland and J.-P. Bouchaud. On a multi-timescale statistical feedback model for volatility fluctuations. *arXiv preprint physics/0507073*, 2005.

J. Boudoukh, M. P. Richardson, and R. Whitelaw. A tale of three schools: Insights on autocorrelations of short-horizon stock returns. *Review of financial studies*, 7(3):539–573, 1994.

D. Challet, R. Chicheportiche, M. Lallouache, and S. Kassibrakis. Statistically validated lead-lag networks and inventory prediction in the foreign exchange market. *Advances in Complex Systems*, page 1850019, 2018.

R. Chicheportiche and J.-P. Bouchaud. The fine-structure of volatility feedback I: Multi-scale self-reflexivity. *Physica A: Statistical Mechanics and its Applications*, 410:174–195, 2014.

P. K. Clark. A subordinated stochastic process model with finite variance for speculative prices. *Econometrica*, 41(1):135–155, 1973. ISSN 00129682, 14680262. URL `http://www.jstor.org/stable/1913889`.

M. Cordi, D. Challet, and I. M. Toke. Testing the causality of Hawkes processes with time reversal. *Journal of Statistical Mechanics: Theory and Experiment*, 2018(3):033408, 2018.

M. Dacorogna, U. Müller, O. Pictet, and R. Olsen. Modelling short-term volatility with GARCH and HARCH models. 1998.

S. Ghashghaie, W. Breymann, J. Peinke, P. Talkner, and Y. Dodge. Turbulent cascades in foreign exchange markets. *Nature*, 381(6585):767, 1996.

L. Gillemot, J. D. Farmer, and F. Lillo. There's more to volatility than volume. *Quantitative Finance*, 6(5):371–384, 2006.

C. H. Hommes. Heterogeneous agent models in economics and finance. *Handbook of computational economics*, 2:1109–1186, 2006.

N. Jegadeesh and S. Titman. Overreaction, delayed reaction, and contrarian profits. *The Review of Financial Studies*, 8(4):973–993, 1995.

D. Kroujiline, M. Gusev, D. Ushanov, S. V. Sharov, and B. Govorkov. Forecasting stock market returns over multiple time horizons. *Quantitative Finance*, 16(11):1695–1712, 2016.

A. Lancichinetti and S. Fortunato. Community detection algorithms: a comparative analysis. *Physical review E*, 80(5):056117, 2009.

M.-X. Li, V. Palchykov, Z.-Q. Jiang, K. Kaski, J. Kertész, S. Miccichè, M. Tumminello, W.-X. Zhou, and R. N. Mantegna. Statistically validated mobile communication networks: the evolution of motifs in european and chinese data. *New Journal of Physics*, 16(8):083038, 2014.

T. Lux et al. Turbulence in financial markets: the surprising explanatory power of simple cascade models. *Quantitative finance*, 1(6):632–640, 2001.

P. E. Lynch, G. O. Zumbach, et al. Market heterogeneities and the causal structure of volatility. *Quantitative Finance*, 3(4):320–331, 2003.

M. Marsili and M. Piai. Colored minority games. *Physica A: Statistical Mechanics and its Applications*, 310(1-2):234–244, 2002.

G. Mosetti, D. Challet, and Y.-C. Zhang. Minority games with heterogeneous timescales. *Physica A: Statistical Mechanics and its Applications*, 365(2):529–542, 2006.

U. A. Müller, M. M. Dacorogna, R. D. Davé, O. V. Pictet, R. B. Olsen, and J. R. Ward. Fractals and intrinsic time: A challenge to econometricians. *Unpublished manuscript, Olsen & Associates, Zürich*, 1993.

U. A. Müller, M. M. Dacorogna, R. D. Davé, R. B. Olsen, O. V. Pictet, and J. E. Von Weizsäcker. Volatilities of different time resolutions—analyzing the dynamics of market components. *Journal of Empirical Finance*, 4(2-3):213–239, 1997.

F. Musciotto, L. Marotta, J. Piilo, and R. N. Mantegna. Long-term ecology of investors in a financial market. *Palgrave Communications*, 4(1):92, 2018.

M. Rosvall and C. T. Bergstrom. Maps of random walks on complex networks reveal community structure. *Proceedings of the National Academy of Sciences*, 105(4):1118–1123, 2008.

M. Tumminello, S. Micciche, F. Lillo, J. Piilo, and R. N. Mantegna. Statistically validated networks in bipartite complex systems. *PloS one*, 6(3):e17994, 2011.

M. Tumminello, F. Lillo, J. Piilo, and R. N. Mantegna. Identification of clusters of investors from their real trading activity in a financial market. *New Journal of Physics*, 14(1):013041, 2012.

G. Zumbach. Time reversal invariance in finance. *Quantitative Finance*, 9(5):505–515, 2009.

G. Zumbach and P. Lynch. Heterogeneous volatility cascade in financial markets. *Physica A: Statistical Mechanics and its Applications*, 298(3-4):521–529, 2001.