Time-dependent relations between gaps and returns in a Bitcoin order book

ROBERTO MOTA-NAVARRO †‡*‡, PAULINO MONROY-CASTILLERO ‡§¶ and FRANCOIS LEYVRAZ †‡

†Instituto de Ciencias Físicas, Universidad Nacional Autónoma de México, Cuernavaca, Morelos 62210, México
‡Centro Internacional de Ciencias AC, 62131, Cuernavaca, Morelos 62131, México
§Department of Solar Energy and Environmental Physics, Swiss Institute for Dryland Environmental and Energy Research, Blaustein Institutes for Desert Research (BIDR), Ben-Gurion University of the Negev, Sede Boqer Campus, 8499000 Midreshet Ben-Gurion, Israel
¶General Coordination of Government Innovation, Government of Jalisco, Ciudad Creativa Digital, Guadalajara 44100, Mexico

(Received 2 July 2021; accepted 11 February 2022; published online 5 April 2022)

Several studies have shown that large changes in the returns of an asset are associated with the sized of the gaps present in the order book. In general, these associations have been studied without explicitly considering the dynamics of either gaps or returns. Here we present a study of these relationships. Our results suggest that the causal relationship between gaps and returns is limited to instantaneous causation.

Keywords: Causal discovery methods; Order book gaps; Large price changes

JEL Classifications: C01, C12, C14, C22

1. Introduction

The study of financial markets has revealed that some of the properties of the price formation process can be explained to a great extent as a consequence of the market structures upon which trading is conducted rather than as a consequence of the trading strategies employed by the market participants (Biais et al. 1995, Goldstein and Kavajecz 2000, Farmer et al. 2005, Gould et al. 2013). One such example are continuous double auction markets, in which trades are mediated via a limit order book.

Several studies have shown that the statistical properties of the orders stored in a limit order book have an effect on price formation. For example, an agent-based model proposed by Farmer et al. (2005) is able to predict 96% of the variance in the spread of 11 stocks, even though the model has only one free parameter and assumes no rationality-based strategies in the agents, showing that the structure of an order book alone plays a big role in the price formation process. Goldstein and Kavajecz (2000) report that as a consequence of decreasing the minimum price variation (tick size) between orders in the NYSE, going from 1/8 to 1/16 of a dollar, the spreads and volumes reduced their sizes, favoring thus participants which consume liquidity from the book.

Other studies have shown that there are statistical relationships between the density of price levels occupied by limit orders in the book and the price fluctuations. In the agent-based models presented by Cristelli et al. (2010) and Schmitt et al. (2012), it was shown that when large gaps between occupied large levels are present in the book, the tails of the returns distribution become fat, so that large changes in price are more common. Of particular relevance to the investigation of the relationships between order book structure and price dynamics is the empirical study conducted by Farmer et al. (2004), which showed that contrary to the intuitive assumption of large price fluctuations being a consequence of orders with large volumes arriving to the book (Gabaix et al. 2003) and consuming several occupied price levels in a single event, the largest fluctuations are actually caused by the presence of large gaps in the order book.

These studies suggest that there is a relationship resembling causality between the gaps in the order book and the returns in the sense that large price fluctuations seem to appear only when large gaps are present in the book. To reach this conclusion, the aforementioned works rely on comparisons among
members of a population of order books and in the observation of the properties of the returns and gaps collected over time, without tracking the changes in gap sizes over time and the subsequent changes in returns they would induce.

Thus it is natural to wonder whether, if, in addition to the results obtained at a population level already reported in the literature, one can observe the same causal relationship in a single instance of an order book as the evolution of the gaps is tracked over time. To help us answer this question, we make use of direct causal discovery methods that are designed to take into account the chronological order of the states of the gaps and returns so that we are able to determine if a cause and effect correspondence exists in this scenario. The confirmation and quantification of causal relationships between the present state of the gaps and future states of the returns could, in principle, allow us to build improved volatility forecasting models that take into account the granularity of the order book as a regressor variable.

In the following, we will focus our analysis on a data set from the Bitcoin/USD market because extensive order book states were available freely. Our methodology follows that of the aforementioned studies on causal relationships between order book granularity and return sizes with regard to the scope of the independent and dependent variables considered: we focused on empirical causal discovery between the states of the order book as a set of multivariate time series which is then divided into non-adjacent time windows to extract different statistics from each of them, as well as the procedure to obtain the Lagged Correlation average.

The structure of the remaining sections is as follows. In Section 2, we discuss the methods employed to represent the states of the order book as a set of multivariate time series which is then divided into non-adjacent time windows to extract different statistics from each of them, as well as the procedure to obtain the Lagged Correlation average. Then we describe the three causal discovery methods applied: Granger Causality, Additive Noise Models (ANM) and Transfer entropy. Section 3 contains the results of the correlation and causality tests between the different series extracted from the data, and finally we discuss the relevance of these results and conclude in Section 4.

2. Methods

2.1. Data pre-processing

We analyzed a set of order book states from the Bitcoin trading platform BTC-e containing the first 20 price levels of each side of the book at a resolution of 10 seconds per state. The data ranges from January 2015 to August 2016 but only 14 months from this time span were included in our study since we discarded months with missing data. The months discarded are June, July and September of 2015 and January and March of 2016.

Figure 1. Procedure to obtain the series of maximum (median) gaps (returns). We divide the series \( s(t) \) into a disjoint windows and extract from within each of these windows the maximum (median) value of \( s(t) \). In the case of the gaps, the series are multivariate, containing 20 price levels per side of the book.

The raw data of the order book can be described as follows: let \( a^1_t \) (\( b^1_t \)) denote the \( s \)th ask (bid) price discrete time indexed by \( t \), with \( n = 1 \) corresponding to the best ask (bid). \( n \) runs from 1 to 20 in our data set. We then define the gaps

\[
\begin{align*}
g^1_s(t) &= \ln(a^1_{s+1}/a^1_t) \\
g^2_s(t) &= \ln(b^1_{s+1}/b^1_t)
\end{align*}
\]

for \( s \in \{1, 2, \ldots, 19\} \). We then choose a window size \( \tau \) and define the time series of the maximum gaps (i.e. the 100th percentiles) for a given month of data as

\[
g_{100}(k) = \max_{1 \leq t \leq T} g^s_m(t)
\]

where \( k \in \{0, 1, \ldots, \lfloor T/\tau \rfloor - 1\} \) are the temporal indices of the windows, \( T \) is the amount of order book states in the given month of data at the maximum resolution of 10 seconds per state and \( \lfloor \cdot \rfloor \) indicates the floor function. We define \( g_{50}(k) \) as the time series of median gaps (that is, the 50th percentiles) in a similar way. This procedure is illustrated in figure 1.

Throughout this work, we will use the term returns to refer to the absolute logarithmic returns, defined as

\[
r(t) = \ln \left( \frac{p(t)}{p(t-1)} \right)
\]

where \( p(t) = (a^1_t + b^1_t)/2 \) is the middle price at time \( t \). We chose to use middle prices to be consistent with previous works (Farmer et al. 2004, Cristelli et al. 2010) that explore the relationship between gaps and returns.
From these we obtain the series of maximum returns which are defined as

\[ r_{100}(k) = \max_{k: \tau = \lfloor (k+1)T \rfloor} r(t) \tag{4} \]

with a similar definition for the median returns \( r_{50}(k) \).

The reasons behind our choice to split the data into windows are two. The first is that by using the middle price to estimate the returns, a spurious and systematic dependence between the first gaps and the returns is introduced; by utilizing windows we can mitigate this effect to some extent. The second reason is that in a real financial market one would usually not be concerned with estimating the size of the return in the next instant of time, but rather the size of the maximum return within a given future time interval, which is precisely what causal discovery methods are equipped to do if a causal relationship is indeed present in the data.

### 2.2. Gaps-returns lagged correlations

Once the series of the maximum (median) gaps and returns are obtained, we compute the average of correlation functions obtained locally over small periods of time between the gaps and return statistics via the following procedure: the gap and return series \( G(k) \) and \( R(k) \) (here \( G(k) \) stands for either \( g_{100}(k) \) or \( g_{50}(k) \), similarly for \( R(k) \)) are partitioned into \( M \) disjoint windows of length \( \tau \). Next, a local lagged correlation is associated to the \( m \)th window in the returns, which is left fixed, by calculating its cross-correlations with a sequence of windows of length \( \tau \) located on \( G(k) \). Each window on \( G(k) \) has a lag in the index \( k \) relative to the fixed \( m \)th window in \( R(t) \) and is thus identified with its corresponding lag value \( l \in \{-\tau, -\tau+1, \ldots, \tau-1, \tau\} \). In this way, we end up with the local Lagged Correlations

\[ C_l(l) = \frac{1}{\tau \sigma_G \sigma_R} \sum_{i=1}^{\tau+1} (R_i - \bar{R})(G_{i+l} - \bar{G}) \tag{5} \]

associated to the \( m \)th returns window, as shown in figure 2. Here \( \bar{R} \) and \( \bar{G} \) denote the averages and \( \sigma_G \) and \( \sigma_R \) the standard deviations of the returns and gaps within the \( m \)th window respectively. Finally, once we obtain all the local Lagged Correlations associated to each one of the \( M \) windows in the disjoint partition of \( R(k) \), we calculate the average Lagged Correlations between \( G(k) \) and \( R(k) \) as

\[ C(l) = \frac{1}{M} \sum_{m=1}^{M} C_m(l) \tag{6} \]

By averaging over the ensemble of local Lagged Correlations we make sure to highlight the geometrical properties that are present in a majority of the local functions and mitigate fluctuations around the average behavior.

### 2.3. Granger causality tests

It is said that a signal \( X(t) \) Granger-causes another signal \( Y(t) \) if a linear model including both the past of the explained variable \( Y(t) \) and the causing variable \( X(t) \) generates better predictions of the future of \( Y(t) \) than a linear model using only the past of \( Y(t) \). The performance of each model is evaluated by minimizing their sums of squared residuals (SSR) and comparing them, with a smaller quantity indicating better prediction capability of the model. There are various ways in which a Granger causality test can be conducted, the particular implementation used in this work is as follows. A linear autoregressive model with a maximum lag of \( L \) (AR(L)) is fit to \( Y(t) \):

\[ Y_R(t) = c_R + \sum_{l=1}^{L} \alpha_l Y(t-l) + \epsilon(t) \tag{7} \]

This linear regression of the explained variable is called the restricted model.

Another linear regression of \( Y(t) \), called full model, is also fitted, but this time with the inclusion of past realizations of \( X(t) \) as regressors:

\[ Y_F(t) = c_F + \sum_{l=1}^{L} \alpha_l Y(t-l) + \sum_{l=1}^{L} \beta_l X(t-l) + \eta(t) \tag{8} \]

with \( \alpha, \beta, c_F, c_R \) being the fitted coefficients of the models and \( \epsilon(t), \eta(t) \) the error terms. Here, the index \( t \) runs from \( t = 1 \) to \( t = T \).

The hypothesis test formulated to assert whether \( X(t) \) Granger causes \( Y(t) \) is that the predictions of the full model are not statistically better than those of the restricted model. To determine this, a standard F-test is usually performed, with the statistic

\[ s = (SSR_R - SSR_F)/L \]

\[ SSR_F/(T - 2L - 1) \]

which follows, under a certain set of assumptions, an \( F_{L,T-2L-1} \) distribution, thus allowing one to compute the probability of having observed a given difference in variances between both linear models.

Usually one finishes a Granger Causality test with an affirmation (or negation) of the causal relationship between \( X(t) \) and \( Y(t) \), when the hypothesis was discarded (or not), under a certain critical value \( \alpha \). In this work, we rather based our conclusions directly on the statistic \( s \) since this number serves as a proxy for the size of the effect that the inclusion of \( X(t) \) has on the predictability of \( Y(t) \), beyond merely answering in a binary way whether the presence of such a relationship is statistically defensible. The reason to proceed in this manner is that the distribution observed for \( s \) differs from the expected F-distribution, which is unsurprising, since both the returns and gaps display fat tails. To make the effect sizes easier to interpret, we define them simply as

\[ s = (SSR_R - SSR_F)/SSRF \]

where we do not normalize by the degrees of freedom of the F distribution corresponding to the test statistic \( s \), such modification does not alter the qualitative aspects of the results since we compare them with adequate controls. We will present the results of the Granger Causality tests as a histogram of the \( s \) values as they vary from one time window to the next and
compare this with surrogate data generated through a random shuffle of the original data, so that it does not show causal behavior.

We will use the nomenclature standard Granger to refer to the tests that only take into consideration lagged versions of the suspect causal variable and instantaneous Granger for the tests that include the causal variable with zero lag, as described in Kirchgässner and Wolters (2007).

The instantaneous Granger Causality tests are important because they will allow us to better distinguish whether the effects are essentially instantaneous or whether they can be used for forecasting.

2.4. Additive noise model causal discovery

Causal discovery through Additive Noise Models (ANM) (Mooij et al. 2016) is a technique in which the explained variable $Y$ is modeled as a function of the causing variable $X$ with an additive and independent noise term $\epsilon$ that acts as a proxy of the cumulative effect of every other latent causal variables that could affect the explained variable. So, the explained variable $Y$ is modeled as

$$Y = f(X) + \epsilon.$$  \hfill (11)

The central idea behind this technique is to assume that the data from two time series $\{x_i, y_i\}, i = 1, \ldots, T$ are variates of a pair of random variables $X$ and $Y$, and that they are related via a one-way causal relationship $X \rightarrow Y$ in which $X$ directly affects $Y$ without the inverse relationship existing at the same time (that is, $Y$ affecting $X$) and assuming that any other causal effect that could come from causal variables other than $X$ can be represented as additive noise $\epsilon$, independent of $X$, as shown in equation (11).

To conduct a causality test by ANM, one first needs to obtain an estimation of the function $f$ which we will denote $\hat{f}$ via a suitable regression method and then, from $\hat{f}$, obtain the residuals $\hat{\epsilon}_i$ as

$$\hat{\epsilon}_i = y_i - \hat{f}(x_i).$$  \hfill (12)

After this is done, one tests the assumption that the variable $X$ is independent of $\epsilon$. This can be done, given the regression analysis, by looking for a correlation between $x_i$ and the residuals $\hat{\epsilon}_i$.

In this work, we made use of the Python package Causal Discovery Toolbox (CDT) (Kalainathan et al. 2020) which includes a module for ANM discovery. The CDT implementation uses Gaussian Process Regression (Roberts et al. 2013) to perform the regression analysis. Here, a Gaussian process is a defined as a multivariate Gaussian probability distribution $p(f)$ over possible functions $f$ that best fit a set of points. Before the hyperparameters that best fit the data are estimated, functional forms for the mean vector $\mu$ and covariance matrix $\Sigma$ of the multivariate gaussian distribution $p(f)$ have to be specified (Görtler et al. 2019). In particular, the covariance matrix is defined in terms of a function $k(x_i, x_j)$ called the kernel so that

$$\Sigma = \begin{bmatrix} k(x_1, x_1) & \ldots & k(x_1, x_N) \\ \vdots & \ddots & \vdots \\ k(x_N, x_1) & \ldots & k(x_N, x_N) \end{bmatrix}$$  \hfill (13)

In the computations performed by the CDT package, the chosen kernel is a radial basis function, defined as

$$k(x_i, x_j) = \sigma^2 \exp\left(-\frac{|x_i-x_j|^2}{2\sigma^2}\right).$$  \hfill (14)

And the mean vector $\mu$ is set to 0.

With respect to the test of independence between $X$ and $\epsilon$, a standard statistical dependence test is used, namely, the Hilbert Schmidt Independence Criterion (HSIC) (Gretton et al. 2005). Let us denote the magnitude of the HSIC statistic by $Z_{X \rightarrow Y} \geq 0$, this serves as an estimation of the possibly non-linear correlation between $x_i$ and the residuals $\hat{\epsilon}_i$. If $Z_{X \rightarrow Y}$ is big enough so that a correlation is detected, then it is discarded that $X$ and $Y$ follow an additive noise model in the direction $X \rightarrow Y$.

With these definitions at hand, we now denote the effect size of the ANM causal discovery method as the score:

$$S_{X \rightarrow Y} = Z_{Y \rightarrow X} - Z_{X \rightarrow Y}$$  \hfill (15)

which is also provided by the CDT implementation. The advantage of using the score $S_{X \rightarrow Y}$ over $Z_{X \rightarrow Y}$ is that to validate an ANM in the direction $X \rightarrow Y$ it is not enough to obtain a low value of $Z_{X \rightarrow Y}$, but also a high value of $Z_{Y \rightarrow X}$ to be...
able to discard the causal relationship in the opposite direction. Indeed, merely having a low value of \( Z_{X \rightarrow Y} \), does not distinguish between \( X \) causing \( Y \) and \( X \) being independent from \( Y \). In this way, the sign of \( S_{X \rightarrow Y} \) indicates the direction of the causal relationship if such a relationship is detected.

2.5. Transfer Entropy

Transfer Entropy is a model-free causal discovery method that measures the reduction in the uncertainty about forecasts of future values of a random process \( Y_t \) when one has access to past values of another random process \( X_t \). The reduction in uncertainty is computed via mutual information, which is an information theoretic measure of statistical dependence. The mutual information between two variables \( X \) and \( Y \) is defined as

\[
I(X, Y) = H(X) + H(Y) - H(X, Y)
\]

where \( H(X) \) is the entropy of \( X \) and \( H(X, Y) \) the joint entropy between \( X \) and \( Y \). The definition of Transfer Entropy with lag \( k \) between the processes \( X_t \) and \( Y_t \) is then defined as

\[
T_{X \rightarrow Y}(t) = I(Y_t, X_{t-k} | Y_{t-k}) = H(Y_t | Y_{t-k}) - H(Y_t | X_{t-k}, Y_{t-k})
\]

where \( I(X, Y | Z) \) is the mutual information between \( X \) and \( Y \) conditioned on \( Z \).

Defined this way, Transfer Entropy can be interpreted (Bossomaier et al. 2016) as the degree of uncertainty about the current value of \( Y \) resolved by the past of \( X \), over and above the degree of uncertainty already resolved by the past of \( Y \) alone.

In a similar fashion to the methodology followed with Granger Causality and Additive Noise Models, we will use the statistic \( T_{X \rightarrow Y}(t) \) as a direct measure of causal effect size.

We made use of the PyCausality python package for the implementation of Transfer Entropy tests.

3. Results and discussion

We begin this section with a summary of basic descriptive statistics of the maximum and median gaps and returns.

Figure 3 shows the distributions of the maximum and median gap sizes within disjoint time intervals of 5 minutes. The median gap distribution is heavily quantized as a consequence of the minimum price change of 0.01 USD. Data corresponding to July 2016.

Figure 4 shows the distributions of maximum and median return sizes within disjoint time intervals of 5 minutes. Data corresponding to July 2016.

Table 1. Monthly average statistics for the 15 available months of data at a scale of 5 minutes. The returns distributions considerable greater deviations from Gaussian behavior in relation to the gaps.

|          | Mean       | Std dev   | Skewness | Kurtosis |
|----------|------------|-----------|----------|----------|
| \( r_{100} \) | 1e – 3     | 1e – 3    | 3.86     | 52.91    |
| \( r_{50} \)  | 3.9e – 5   | 1e – 3    | 6.28     | 66.54    |
| \( r_{100} \) | 2.9e – 3   | 1.6e – 3  | 1.93     | 7.52     |
| \( r_{50} \)  | 1e – 4     | 4.9e – 5  | 1.15     | 1.36     |

For the computation of the Lagged Correlations, we considered two different scenarios: either we limited the set of gaps from which we extracted the maximum and medians to from each of these windows the 50th and 100th percentiles (median and maximum), as described by equation (2). Once we derived the series of maximums and medians at the selected scale, we proceeded to calculate the average Lagged Correlations between them using a sliding window of 100 data points, where each data point is one of the maximums or medians extracted previously at a scale of 60 time steps. This procedure was executed on each month of data separately. The average local Lagged Correlations between the percentile series was then computed. Figures 5 and 6 show the Lagged Correlations between the complete set of gaps and the returns and between only the first gaps and the returns, respectively. Note that the magnitudes of the Lagged Correlations are greatest when considering the maximum events and quickly decrease in size as one considers lower percentiles, in particular, the Lagged Correlations between the medians (50th percentile) are very small, compared to those between the maximums.

We measured the Lagged Correlations between the maximum and median gaps and returns by splitting the order book multivariate series into disjoint windows of 60 time steps at a resolution of 10 seconds per time step. Next, we extracted the corresponding range of sizes in a more homogeneous manner.
Figure 5. Average Lagged Correlations between the maximum \((q = 100)\), 99th percentile \((q = 99)\) and median \((q = 50)\) of the full gaps and returns. There is an abrupt change in the magnitude of the Lagged Correlations as one considers percentiles other than the maximum. At a lag of zero, the correlations of 99th percentiles are almost four times smaller than those between the maximums.

Figure 6. When only the first gaps \((g_{11}^1\) and \(g_{12}^1\)) are considered, the Lagged Correlations between the maximums, 99th percentiles and median gaps and returns decrease gradually as lower percentiles are taken into consideration in sharp contrast with the Lagged Correlations of the full gaps.

just the gaps in the first position, or we extracted them from the complete set of available gaps. When only the first gaps are considered \((g_{11}^1\) and \(g_{12}^1\)), we observe that the maximum correlation between any pair of percentiles is attained at zero lag, having lower values for negative and positive lags. When the complete set of gaps is used, on the other hand, the 100th and 99th percentiles (the maximums) attain their maximum correlation at lag zero but the median (percentile 50) has its maximum at positive lags. Since the returns are the ones being fixed, the delayed position of the maximum Lagged Correlations in the median implies that return statistics in the present are associated with future gap statistics. The reasons behind the delayed position of the maximum are unclear but they do not entail a causal relationship between gaps and returns.

The median size of gaps does not correlate as strongly to the median size of returns as do the maximums. By contrast, the Lagged Correlations between the median returns and the maximum gaps are not as weak, as shown in figure 7, thus, both the typical and maximum size of returns are statistically associated with the maximum gaps.

These results also hold for window sizes different from 100 time steps.

3.2. Granger causality tests

The fact that the Lagged Correlations do not vanish immediately for lags different from zero is compatible with a causal relationship but insufficient to demonstrate it. A direct causality measure such as Granger Causality can be used to distinguish whether the relationships observed are consistent with the idea of causal effects or are better explained as correlations.

For this purpose, aside from the lagged correlations between the gap and return percentiles, we conducted simple Granger Causality tests in both causal directions: gaps to returns and vice versa. The Granger tests were conducted at four different timescales \((\tau = 30, 60, 90\) or 120 steps at 10 seconds per step) by dividing each month of the series \(r_{50}, r_{100}, g_{50}\) and \(g_{100}\) into disjoint windows with a length of \(\bar{T} = 500\) time steps, as described in figure 8. This segmentation has the purpose of increasing the sample size of causality scores, which would be too small if we only conducted the tests on the 14 months of available data. It is important to remark that within the series \(r_{50}, r_{100}, g_{50}\) and \(g_{100}\), each time step corresponds to either a maximum or a median gap (return) taken over a scale \(\tau\) as mentioned before.

The maximum lag over the windows (indexed by \(k\)) was chosen via the Bayesian information criterion (Gordon 2015) (BIC), which was performed on the auto-regressive models of each series with lag values ranging from \(l = 1\) to \(l = 10\). We chose the value of \(l\) that achieved the minimum median BIC across this range of lags and across all values of \(\tau\).
Figure 8. Illustration of the partitioning of each month of data into windows to increase the sample size of the Granger tests. A given month (January in the figure) of a series obtained as described in 1 is partitioned into disjoint windows of $\tilde{\tau} = 500$ time steps and a Granger test is conducted between each pair of windows. This procedure is performed on all 14 months of data available and the results are collected to compute the various distributions of Granger Causality scores of figures 9 and 10.

Figure 9. Effect sizes for the Granger Causality tests. Red corresponds to the tests conducted on the original data and blue to the tests on shuffled surrogates. The dotted black line marks the critical value $\alpha = 0.01$. Since the random shuffling destroys any time dependence between time series, the shuffled surrogates represent the null hypothesis of no causality between gaps and returns. The distributions of effect sizes of the tests conducted on the real data are very similar to those conducted on the surrogates both in the shape and support of their distributions, in accordance to a lack of a Granger causal relationship between these two signals.

The effect sizes, corresponding to the non-normalized $s$ statistic mentioned previously (see equation (10)), are shown in the figure 9, where the red histograms correspond to the results of the tests on the original data while the blue ones to the tests of randomly shuffled copies of the data, used as controls. Except for relationship between the median returns and gaps, the distributions of causality scores of the controls and the real data have very similar shapes and supports and very few tests achieved statistical significance under a critical value of $\alpha = 0.01$, with only 12% of tests surpassing $\alpha$ in the case of the relationship $g_{100} \rightarrow r_{100}$ and 10% in the case of $r_{100} \rightarrow g_{100}$.

In the standard Granger tests as well as in the ones that include instantaneous effects, no asymmetry between the effect sizes of both causal directions $g_{100} \rightarrow r_{100}$ and $r_{100} \rightarrow g_{100}$ is evident, as such, it is roughly equally likely to measure a given effect size in either direction.

When instantaneous windows of the causal variable are added in the linear regressions of the explained variables, that is, when the minimum lag in the regressive models is zero, a very different result is observed, as shown in figure 10. Although the controls and the real data distributions are similar for the medians, for the maximums the shapes of the distributions become very dissimilar and the effect sizes blow up, with 100% of the tests achieving statistical significance in both causal directions (table 2).

As we mentioned before, the distribution of the Granger Causality statistic score $s$ differs from the expected F-distribution as a consequence of the fat tails present in both the returns and gaps distributions. The reader might wonder why then it is justified to still use this causal discovery technique. The answer to this issue lies in the distinguishability between the distribution of the statistic under the null hypothesis of independence (as is the case with surrogate data created through random shuffles) and the distribution obtained when a
Table 2. Proportions of statistically significant causality scores in the Granger tests for the null hypothesis of no causal effect under a critical value of $\alpha = 0.01$. The standard tests throw a very small number of significant tests. When we consider the inclusion of instantaneous windows in the linear models of the test, the causality scores blow up, with every single test surpassing the critical region, in the case of the maximum gaps/returns. The medians remain with low percentages of significant tests.

|                      | Standard | Instant |
|----------------------|----------|---------|
| $g_{100} \rightarrow r_{100}$ | 0.12     | 1.00    |
| $r_{100} \rightarrow g_{100}$ | 0.10     | 1.00    |
| $g_{50} \rightarrow r_{50}$ | 0.04     | 0.11    |
| $r_{50} \rightarrow g_{50}$ | 0.30     | 0.27    |

Table 3. Proportions of statistically significant causality scores in the ANM tests at a level of $\alpha = 0.01$. We see that when the gaps and returns series are taken without a lag a considerable proportion of tests result statistically significant.

|         | $L = 0$ | $L = 1$ | $L = 3$ | $L = 7$ |
|---------|---------|---------|---------|---------|
| $g_{100} \rightarrow r_{100}$ | 0.35 | 0.12 | 0.05 | 0.03 |
| $r_{100} \rightarrow g_{100}$ | 0.21 | 0.26 | 0.08 | 0.03 |

causal effect is present. We show in Appendix that even when the data is heavy tailed, the distributions of $s$ statistics are clearly distinguishable so that a non-parametric hypothesis as conducted in this work becomes a valid procedure.

3.3. Additive noise models tests

Now we describe the results of the application of causal discovery by means of Additive Noise Models. Since ANM tests do not inherently consider the chronological order between the causal and explained variables as Granger Causality does, we performed tests between the gap and return percentile series considering no lag between them, which corresponds to instantaneous causal relationships, and between lagged versions of the series, mimicking the chronological order taken in the standard Granger Causality tests. There are two key differences between this way of forcing chronological order into the ANM tests and the way in which Granger Causality naturally considers time via the lags in its regressive models. The first difference is that for a given maximum lag value $L \in \mathbb{N}$, the test of Granger takes into account lagged observations ranging from 1 to $L$ time-steps in the past while the ANM test will only consider a lag of $L$ time-steps. The second difference is that in the case of non-zero lags, only the positive results have meaning since negative causality scores would correspond to causal effects from the series located in the future to the series located in the past.

The results of the ANM tests, shown in figures 11, 12, 13 and table 3, are in agreement with the Granger Causality tests in both cases: the case where no lag is considered, which shows the highest deviation of causality scores with respect to the control scores distribution, and the lagged cases, where the distribution of scores becomes similar to the control distribution, even when considering a lag of a single time step.

Figure 11. Effect sizes for the ANM tests applied to the maximum returns and gaps with a lag value of $L = 0$. A positive effect size ($S_{g_{100},r_{100}}$) corresponds to a causal effect from the gaps size to the returns size and a negative one ($-S_{r_{100},g_{100}}$) to an effect from the returns to the gaps. The dotted black lines mark the two-tailed critical at $\alpha = 0.01$. The shapes of the control and data distributions are very dissimilar and a significant proportion of the cases fall far away from the tails of the controls.

Figure 12. Effect sizes for the ANM tests for the causal relationship $g_{100} \rightarrow r_{100}$ with positive lag values. Only the positive causality scores correspond to a detection of causal effects since negative values would imply that the future return sizes have an effect on past gap sizes. The dotted black lines mark the one-tailed critical values at $\alpha = 0.01$. As opposed to the case with zero lag, now the shapes of the control and data distributions are similar and only a small fraction of the tests fall far from the tails of the controls.
For the causal relationships between the medians, no statistically significant effect is found independently of the value of the lag since the ANM causality scores are totally contained within the values of the control distribution. A peculiarity of the causality scores between the medians is that these are all negative for the data and the controls. A possible explanation of this phenomenon is given in the discussion section.

### 3.4. Transfer entropy tests

In a similar fashion to the model underlying the Granger Causality test, the method of Transfer Entropy inherently takes into consideration time lagged versions of the causal variables with respect to the explained variable, as shown in equation (17). With the difference being that Transfer Entropy considers only a single lag.

The results of the instantaneous causality Transfer Entropy tests are shown in figure 14 and the results of the standard, time-lagged tests, in figure 15.

The entire sample of tests surpasses the critical value of the control distribution for the instantaneous causal relationships $r_{100} \rightarrow g_{100}$ and $g_{100} \rightarrow r_{100}$. On the other hand, a vast majority of the tests reside within the limits of the control distribution for the relations $r_{50} \rightarrow g_{50}$ and $g_{50} \rightarrow g_{50}$.

The results with a nonzero lag $L = 1$ are shown in figure 15 and confirm what was obtained with the Granger and ANM methods. Only small fractions of the tests achieved statistical significance.

We repeated the Granger, ANM and Transfer Entropy tests between the gaps and two different definitions of volatility: as
the average of absolute logarithmic returns and as the standard deviation of logarithmic returns, obtaining the same qualitative results. Further, we conducted standard Granger Causality tests were instead of taking the maximum gap within each disjoint time window, we took the maximum sum of the first three gaps. By doing this we consider the aggregated effect of the maximum gap, which is usually located within the first three positions, along with the effect of the next two biggest gaps, which are also usually located within the first three positions. We have not displayed these numerical results to save space, as they are not significantly different from the other results discussed in the text (table 4).

4. Discussion

4.1. Simultaneous causal relationships

There are statistical associations between the maximum gaps and maximum returns and the maximum gaps and median returns that do not immediately die off as evidenced by the correlation functions shown in figures 6 and 7. With the exception of the median gaps and returns, these correlations attain their highest magnitude at zero lag in a similar way to that in which the causality tests attain the highest scores when the observations include instantaneous windows.

In such scenario, all the tests for Granger Causality and Transfer Entropy become statistically significant in both causal directions: *g*_{100} → *r*_{100} and *r*_{100} → *g*_{100}, which only indicate a strong simultaneous relationship between maximum gaps and maximum returns but without pointing out a particular direction (figure 10), this indicates that there is no direction to the ‘causality’ observed. Furthermore, the ANM tests differ significantly at both tails of the causality scores distribution with respect to the control distribution (figure 11).

So, with the inclusion of instantaneous time periods, the Granger and Transfer Entropy tests fully reject the null hypothesis of no causal effect. Similarly, although the ANM tests do not fall completely outside of the control distribution support, they make it difficult to accept the null hypothesis since more than half of the tests reach statistical significance with a difference of 1.66 times more tests favoring the direction *g*_{100} → *r*_{100} than the opposite direction *r*_{100} → *g*_{100}.

These results imply that there is a link between the size of the maximum gap and maximum return of mid-price within the same time interval \([t, t + \tau]\). Therefore, the correlation functions and the causality tests with instantaneous windows are consistent with the results shown in previous literature (Farmer et al. 2004, Cristelli et al. 2010) since they show that the existence of large gaps in the book is associated with the generation of large returns, when returns are defined as mid-price returns.

There are, however, caveats with the positive results of the instantaneous causality tests: since the returns were defined as a function of the middle price, an artificial dependence of the returns on the first gaps of the order book is induced. The maximum gap is usually the first one: in the data analyzed around 30% of the maximum gaps are located at the first level, as shown in figure 16. Appearance or disappearance of solely a first gap due to any reason (limit orders placements, cancellations or transactions) implies a change of mid-price proportional to the size of that first gap. Thus, maximum returns within the time interval \([t, t + \tau]\) often correspond to changes when the first gap is also the maximum gap. This correspondence naturally rises a strong statistical relation between the sizes of the maximum returns and the first gaps.

For these reasons, we must be very careful when interpreting the positive results of the instantaneous causality tests.

4.2. Time lagged causal relationships

With respect to the standard causal scenario where lagged versions of the observations are considered the Granger, Transfer Entropy and ANM tests throw mostly non-significant results (figures 9, 15, 12, 13 and table 3). Among all tests, only few of them achieve statistical significance, so their distributions for each case are very similar to their respective control distributions. This similarity is important since randomly permuted surrogates keep the same moments and distribution as the original data but any structure that depends on the temporal

Table 4. Proportions of statistically significant causality scores in the Transfer Entropy tests under a critical value of \(\alpha = 0.01\). Most tests result in non statistically significant causality scores when a lag of \(L = 1\) time steps is used, either between the maximum or the median gaps and returns. When we force the inclusion of instantaneous terms, every test with the maximum gaps/returns is significant while the medians remain mostly non significant.

| Direction   | Standard | Instant |
|-------------|----------|---------|
| *g*_{100} → *r*_{100} | 0.012 | 1.00    |
| *r*_{100} → *g*_{100} | 0.002 | 1.00    |
| *g*_{50} → *r*_{50}  | 0.008 | 0.11    |
| *r*_{50} → *g*_{50}  | 0.012 | 0.27    |

Figure 16. Distribution of the positions at which the maximum gaps is located. The sign of the positions correspond to the sides of the book: positive for the asks and negative for the bids. That is, a value of \(-1\) indicates the first gap in the bids and correspondingly, a value of 1 indicates the first gap in the asks. It is unlikely to find the maximum gap beyond the fifth position, with a probability of roughly 30% of being located in the first position of either side.
order is destroyed (Lancaster et al. 2018), which turns them into a very aggressive control scenario. So, even when we destroy all chronological structure we get very similar causality score distributions to the ones obtained in the original data, a result that strongly suggests the absence of a cause and effect relationship.

As mentioned earlier, all the ANM tests for the medians resulted non statistically significant since the support of the distribution of causality scores from the data is completely contained within the support of the controls distribution, with the peculiarity that in all tests, both in the data and the controls, the test statistic is negative. This would suggest the causal direction $r_{50} \rightarrow g_{50}$ if the tests were statistically significant, which is not the case. Aside from that, the correlation function between the median gaps and returns shown in figure 5 and the distribution of Granger Causality test effect sizes in the direction $r_{50} \rightarrow g_{50}$ shown in figure 9 could suggest the existence of a small causal effect on the median size of the gaps stemming from the median returns. We think this is not the case since the median gaps usually reside above the first levels in the order book, where the market activity rarely consumes volume. Although we do not have definite answer to this issue, a possible explanation could be that this is an artifact of the quantized nature of the median gap sizes along with the fact that the median gaps explore only a small region of possible values, as opposed to the maximum gaps that explore a wide range of values from a few cents to several dollars, as seen in figure 3.

Taken together, the nullity of the standard causality tests suggest that the size of the largest return found within the time interval $[t + \tau + 1, t + 2 \tau + 1]$ cannot be causally linked to the size of the largest gap found in the preceding interval $[t, t + \tau]$, again, to the extent of the causality definitions assumed by the techniques we used and the time scales we explored.

It does not appear to be the case that the results found previously in the literature explaining the differences in the tails of the returns distributions of different companies as a consequence of the granularity of their respective order books apply when the granularity of a single order book is tracked over time. This could be a particularity the bitcoin/dollar market so a replication of this study on the stock market would be of use to test for this possibility.

5. Conclusion

The work presented here broadens previous works on the relationship between the granularity of the order book and the returns across a population of order books in which we took into consideration a single instance of an order book and tracked the evolution of the gap sizes across time with the aid of causal discovery methods in an attempt to test the hypothesis of the presence of large gaps in the book as a primary source of large returns. We have shown for the Bitcoin/USD trading platform BTC-e, that the evolution across time of large price fluctuations does not seem to be a consequence of fluctuations in the size of the gaps present in the book in the sense that having access to the present state of the gaps appears to be of little use to create estimates of the future size of maximum returns in future windows of time. On the other hand, when the same time interval is considered in both the series of gaps and returns, we found that the size of the maximum gap is an excellent predictor of the size of the maximum return. These results could be a particularity of the financial market that we studied or of the minimum time scales imposed by the data but at the very least they suggest restrictions on the temporal horizon and precise definition of causal relationship that can be expected between the granularity of the book and the tail properties of the returns. Knowing such restrictions can be of help to better guide the sources of information considered in the creation of volatility forecasting models. A replication of our tests on different markets, at smaller timescales and possibly with complementary causal discovery tools would be of much use to obtain stronger constraints on the conclusion that knowledge of the state of the gaps in the order book does not seem to benefit the estimation of future price fluctuations.

Acknowledgements

We would like to thank Hernán Larralde for his helpful discussion and feedback.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

The authors gratefully acknowledge the financial support from CONACyT, Grant Nos. Proyecto Fronteras 201 and CONACyT CB 254515, and the UNAM PAPIIT research, Grant No. IN113620. Roberto Mota Navarro thanks CONACyT for a Ph.D. scholarship.

ORCID

Roberto Mota-Navarro  http://orcid.org/0000-0002-6077-7386
Paulino Monroy-Castillero  http://orcid.org/0000-0002-1193-9991
Francois Leyvraz  http://orcid.org/0000-0001-9269-1248

References

Benzaquen, M., Mastromatteo, I., Eisler, Z. and Bouchaud, J-P., Dissecting cross-impact on stock markets: An empirical analysis. J. Statist. Mech.: Theory Experiment, 2017, 2017(2), 1–21.
Biais, B., Hillion, P. and Spatt, C., An empirical analysis of the limit order book and the order flow in the paris bourse. J. Finance., 1995, 50(5), 1655–1689.
Bossomaier, T., Barnett, L., Harré, M. and Lizier, J.T., An Introduction to Transfer Entropy, 2016 (Springer: Cham, Switzerland).
Cristelli, M., Ali, V., Pietronero, L. and Zaccaria, A., Liquidity crisis, granularity of the order book and price fluctuations. Eur. Phys. J. B, 2010, 73(1), 41–49.
Farmer, J.D., Gillemot, L., Lillo, F., Szabolcs, M. and Sen, A., What really causes large price changes? Quant. Finance, 2004, 4(4), 383–397.
Appendix

In this appendix, we show that when we deal with time series that show heavy tailed distributions, it is possible to perform a reliable non-parametric Granger Causality test, even though the test in its standard parametric version is designed to deal with non-heavy tailed data.

To this end, we performed a numerical experiment consisting in the creation of two pairs of time series $X$ and $Y$ so that in one pair the observations of $X$ follow a Pareto distribution and $Y$ is dependent on past values of $X$ plus an additive noise term, so that

\[ X \sim f(x; b) = \frac{b}{x^{b+1}}, \quad \text{for } x \geq 1, \quad b > 0 \]  

(A1)

and

\[ Y(t) = \alpha X(t - 1) + \epsilon, \quad \alpha \in [0, 1) \]  

(A2)

where the additive noise term $\epsilon$ is normally distributed. And in the other pairs the two series follow a Pareto distribution but are independent of each other.

Figure A1 shows the results of a realization of this numerical experiment for a given set of parameter values. It can be seen there how different the supports and shapes of the causality scores distributions are when no dependence exists between the other pairs the two series follow a Pareto distribution but are independent of each other.

This distinguishability validates the use of the control distribution generated with the randomly shuffled surrogate samples as a non-parametric distribution of causality scores under the null hypothesis, with the added benefit of inherently handling the fat tailed (not necessarily Pareto) nature of the original data.