Market sentiment and inhomogeneous information flow : a cryptocurrency exchange case-study

Joanne Affolter, João Luis Prado Vieira
School of Communication Systems, EPFL, Switzerland

***Abstract*—Cryptocurrency markets are highly dynamic and volatile. They are characterized not only by rapid price fluctuations, but also by their sensitivity to market sentiment shifts relating jointly to the underlying assets and to the exchange platforms where these assets are traded.**

This study examines how market sentiment fluctuations influence changes in currency quotes across different cryptocurrency exchanges, focusing on Bitcoin. We collect tick-by-tick trade data and second-resolution market order book data from Binance and Bitstamp throughout Jan 2021 - Dec 2021 and employ Google Trends search indexes for crypto keywords as proxies. The analysis unfolds in three stages: (i) validating our chosen proxy as an explanatory variable for cryptocurrency price changes using transfer entropy as a measure of information flow, (ii) checking the validity of traditional stylized facts in the crypto case and their variations across both exchanges, and (iii) scrutinizing the time-evolution of the induced lead-lag correlation throughout the studied period.

I. Introduction

Cryptocurrency markets are dynamic and highly volatile systems, whose recent growth is accompanied by the apparition of social networks. Many recent works have aimed to harness market sentiment as a predictor of cryptocurrency returns [1] [2] [3]. However, certain methodologies merely employ proxies of market sentiment as exogenous variables in asset pricing models, neglecting crucial elements of the relationship between market prices and market sentiment, such as the causal relationships, as well as variations across different exchanges, locations and market periods.

In this work, we study the variations in cryptocurrencies price responses to changes in market sentiment across different crypto-exchanges. We focus our analysis on a basket of four cryptoassets, Bitcoin (BTC), Ethereum (ETH), Cardano (ADA) and Litecoin (LTC), and two trading exchanges, Binance and Bitstamp, the largest by market cap and the longest-running exchanges, respectively [4][5]. We concentrate on the timeframe from January 2021 to December 2021, in which all the studied assets attained their maximal historic price. We use intraday Google Trends (GT) search volume indexes for multiple cryptocurrency related keywords as proxies of market sentiment relating to this sector.

The research unfolds through three stages. First, we validate the selected sentiment proxy as an explanatory variable for cryptocurrency volatility, volume and quote price changes. Subsequently, we explore the validity of traditional stylized facts relating to the empirical return distribution, cross-correlation and trade-initiation response function of assets in the cryptocurrency case and scrutinize their variations across both Binance and Bitstamp. Lastly, we examine the time-evolution of the induced lead-lag correlation, unraveling the relationships between market sentiment and exchange-specific dynamics throughout the observed period.

We note that a complete analysis would probably require the addition of novel cryptocurrency products, such as futures, perpetuals and options, since recent work indicates that these asset classes play a role in price discovery [6] and in cross-exchange price synchronicity [7]. However, we exclude cryptocurrency derivatives markets from the present study due to rate limits of our data source and variability of offered products across exchanges. Nevertheless, future work may be interested in analyzing the impact of market sentiment effect in quote prices conditioned on states of options or futures markets.

Related works

Google Trends in finance and econometrics. Launched in 2006, Google Trends (GT) provides location- and time-normalized search interest indexes for Google Search keywords. Google Trend indexes have been used to track economic growth in real-time [8], to infer predictors of market moves [9], or as early indicators of market crashes [9]. However, statistical analysis using GT data faces important technical challenges [10]. First, the risk of data bias and causality breakage stemming from signal availability, since it is not possible to assess if search results associated with a given timeframe were actually available at the given timeframe. Second, the combinatorial untractability of model selection if GT data is used as regressors, since models are parameterized by combinations of keyword choices.

Causal inference and time-asymmetric transfer of information. Statistical studies and tests for causality for multivariate time-series have become increasingly popular in finance, since the advent of Granger causality and cointegration tests [11] [12]. However, application of these methods for real-world data requires unverified or unverifiable properties, such as the need for a linear model of the process dynamics in the case of the former method, and co-integration of same order for all time-series in the latter case. Transfer entropy [13] emerged as a metric of the predictive power of the past of a given predictor time-series X over the future of predicted time-series Y, following a possibly non-linear relation $Y = f(X)$.

IDTxl. IDTxl is a Python tool used for estimating information dynamics such as transfer entropy in data. It uses a greedy algorithm to estimate bivariate transfer entropy (bTE) and multivariate transfer entropy (mTE), also termed conditional

transfer entropy.
Bivariate transfer entropy from one source X to one target Y can be defined as the information the past of X provides about Y in the context of Y's past. It measures in fact the additional information provided by X with respect to the information about the past of Y when one tries to predict Y [14]. Let's denote the future and past of X, Y respectively by X^+, Y^+ and X^-, Y^-. Transfer entropy from X to Y can be written as:

$$TE_{X\to Y} = H(Y^+|Y^-) - H(Y^+|X^-, Y^-),$$

where $H(A|B)$ is the conditional entropy of A on B.
In a multivariate setting with a single target process, Y, and a set of source process $X = \{X_1, ..., X_k\}$, we define the mTE from a source X_i to Y as the information the past of X_i provides about Y, in the context of both Y's past and the past of all other relevant sources in X.

Description of the algorithm used to infer links between sources and targets using either bTE or mTE can be found in appendix VI. Different quantities can be estimated for statistical significant links with IDTxl, such as :

- individual transfer entropy from one source to one target,
- p-values estimated from surrogate data,
- information-transfer delay [15] estimated as the time lag between X's past value and the current value at time n which provides the maximum individual information contribution (indicated by the maximum TE estimate),
- joint information transfer from all source variables into a single target in case of mTE.

Statistical testing of estimated quantities is usually carried out by comparing the original estimate with a distribution derived from surrogate data. To construct this test-distribution, the data is permuted a significant number of times and the quantity of interest is re-estimated. The original estimate is then compared to this distribution, and a p-value is calculated as the proportion of surrogate estimates that exceed the original estimate.

Lead-lag networks. In time-series analysis, the cross-correlation $C_{X,Y}(t,\tau)$ of two time-series X, Y is an important tool to understand their interaction and dynamics. In the asynchronous, discrete case, the empirical estimator of the cross-correlation suffers from drawbacks related to time-series synchronization and sampling. An example of these drawbacks is the Epps effect [16], in which the empirical cross-correlation of two-time series is decreasing with time-lag τ and disappearing at the limit $\tau \to 0$. In order to mitigate this problem, the literature has moved towards more robust asymmetric cross-correlation estimators, such as the Hayashi-Yoshida estimator [17], which mitigates the Epps effect while preserving convergence to the true correlation in the gaussian setting. The asymmetry of the Hayashi-Yoshida estimator is quantified through the Lead-Lag Ratio (LLR), which has been leveraged in the literature as an indicator of causal relations [18] and can be expressed as

$$LLR_{X,Y} = \frac{\rho^2(l_i)}{\sum_{i=1}^{L} \rho^2(-l_i)}$$

where $\rho^2(l_i)$ is the Hayashi-Yoshida estimated cross-correlation between time series X and Y, while the latter is lagged by l_i seconds. We say that "X leads Y", if $LLR_{X,Y} > 1$. The computation of LLRs over a family of assets allows the definition of time-dependent lead-lag networks, which estimate how changes in prices in a given asset influence the price of other assets. More specific to the cryptocurrency sector, recent work has worked on the representation learning of these relations [19] and on the lead-lag duration between cryptocurrencies [20]. To the best of our knowledge, this work is the first studying not only lead-lag networks of different cryptoassets, but also the networks spanned by conditioning trading to a fixed cryptocurrency exchange. We leverage the implementation of the Hayashi-Yoshida estimator from the *Lead-lag* Python library[1].

II. Data collection and processing

A. Historical cryptocurrency data

Data sources. We retrieve daily data from the Yahoo Finance API[2]. Regarding the intraday data, we found a dataset on Kaggle with BTC price data captured at one-minute intervals from the Binance API [3]. No datasets were available for the three other assets. Consequently, historical order book and trade data collected for the last part of the project were used to obtain 1-minute frequency closing prices. To infer closing prices, the approach involved aggregating data by minute, selecting the last transaction, and extracting the mid-price.

Data analysis. One of the main challenges faced during this project was to work with intraday data. Most of the data was collected through API calls, presenting constraints on the daily quota of allowable calls. This prompted an early strategic decision to carefully choose a specific time period for the project's focus. Bitcoin (BTC) naturally emerged as the ideal candidate for pinpointing the most relevant period, given its leading status in the cryptocurrency market. The results of the Bitcoin's analysis are reported in the section IV-A *In-Depth Analysis of Bitcoin*.

B. Google Trends

We employ Google Trends search indexes for a chosen set of keywords as proxies for the market sentiment towards the studied cryptocurrencies, e.g. Bitcoin (BTC), Ethereum (ETH), Cardano (ADA) and Litecoin (LTC). The selected keywords, classified by main topics, are presented in Table I.

The *Google Trends search index*, ranging from 0 to 100, serves as a metric indicating the popularity of a keyword on the Internet. It is derived from the frequency of Google searches related to a specific term. Higher values signify increased interest. To obtain the Google Trends indexes for the studied

[1]Available at https://github.com/philipperemy/lead-lag
[2]Available at https://pypi.org/project/yfinance/
[3]Available at https://www.kaggle.com/datasets/jkraak/bitcoin-price-dataset

Topic	Keywords
Studied cryptocurrencies	Bitcoin, Ethereum, Cardano, Litecoin, BTC, ETH, ADA, LTC
Crypto lexicon	crypto, cryptocurrency, trading
Studied trading exchanges	Binance, Bitstamp
Impacting events	China, FED, Musk

TABLE I: Set of selected keywords.

period, we conducted a series of API calls to the Google Trends API [4].

Technical challenges and improvements. In this work, two main challenges were faced during the collection of Google Trends data.

First, to ensure the acquisition of 1-minute intraday data, GT time-series data points must be collected by slots of 4 hours. Indeed, beyond a specific threshold, the data undergoes aggregation on a 8-minutes, hourly, daily, or even weekly or monthly basis. However, this strategy necessarily implies the need for a considerable number of API calls to collect all data needed for our period of interest. Since Google does not offer a proper API for Google Trends, we were unable to collect 1-minute frequency data as initially planned, due to the low rate of authorized requests (which remains undisclosed, preventing us from tuning our code). We therefore decided to collect hourly data manually instead. Nevertheless, we make the code to perform API requests available in our repository along with demonstration code on the collection of 1-minute frequency GT data.

Secondly, Google Trends does not offer normalized data across time. Values are scaled so that at least one timepoint has a score of 100 within the requested timeframe. Therefore, concatening the results of multiple API calls would give inconsistent results when comparing different time periods.
To overcome this problem, we therefore had to standardize the data. Let's denote the time-series collecting Google Trends indexes for a timeslot i by $\{x_t^{(i)}\}_t$. We decided to scale each time series $\{x_t^{(i)}\}_t$ by a constant factor c_i, calculated as follows :

$$c_i = \frac{x_t^{(i-1)}}{x_{first}^{(i)}},$$

where $x_{first}^{(i)}$ and $x_t^{(i-1)}$ correspond respectively to the first value of the current time-series and to the value in the previous requested time-series with matching index (e.g. day and hour). Note that the timeframes were computed in such a way that two consecutive time-series have one overlapping index. If either one of the value corresponding to the same index is null, the multiplicative factor is set to 1.

After normalization, we would henceforth get :

$$x'^{(i)}_{first} = x_{first}^{(i)} \times c_i = x_{first}^{(i)} \times \frac{x_t^{(i-1)}}{x_{first}^{(i)}} = x_t^{(i-1)}$$

[4]https://trends.google.fr/trends/

and values will be consistent across all the studied timeframe. Fig. 1 shows the evolution of keywords' indexes over the period of interest. The ranking of keywords based on their averaged index is provided in Fig. 3

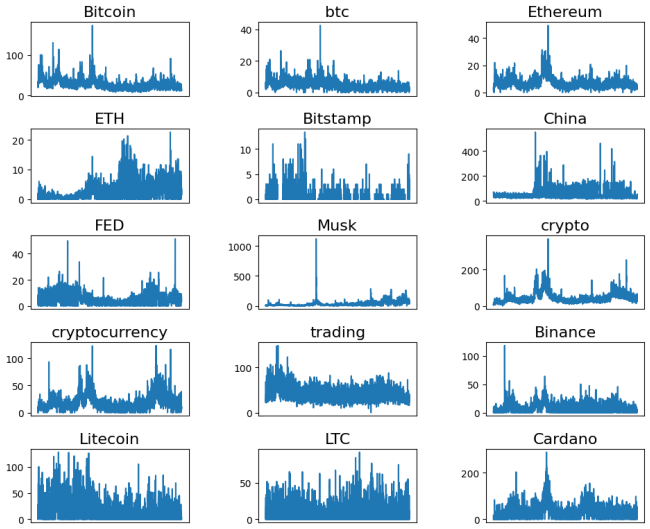


Fig. 1: Google Trends indexes

Fig. 2: Bitcoin's price

As anticipated, some keyword's indexes closely mirror the price dynamics of Bitcoin 2, with an exponentially looking trend up, followed by a U-shaped curve. This trend is particularly evident in the case of "crypto" and "cryptocurrency" keywords. One interesting observation concerns China, where its index suddenly increased at a certain time, a period that seems to be closely aligned with the date of China's crypto ban, as we will see in section IV-A - *In-Depth Analysis of Bitcoin*.

C. Data Pre-Processing of Trade Data

The preprocessing and cleaning pipelines of this work are subdivided in two steps : (i) exploratory data analysis for outlier/artifact detection, time-series alignment, and (ii) partition of the data into sub periods of interest.

During step (i), trading days for which no trades are recorded for periods longer than one hour are excluded from our analysis. We assume these silent periods correspond to moments of platform unavailability and are therefore not representative of the period's real trading volume. We note that opposed to traditional markets, cryptocurrency markets

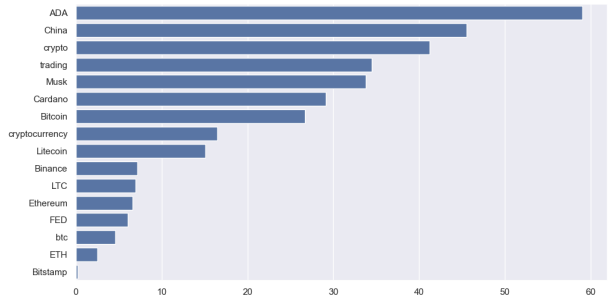


Fig. 3: Ranking of keywords based on their averaged Google Trends index over time

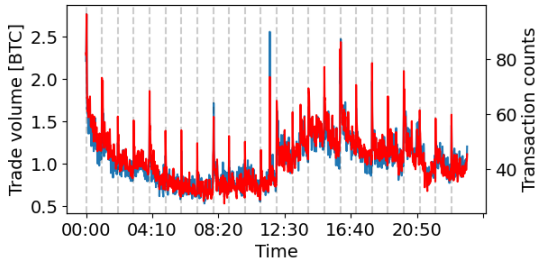


Fig. 4: Mean trade volume (red) and transaction counts (blue) for BTC across a trade day in Binance.

do not have interruptions in trading hours and may present or not the day of the week effect [21] or stylized intraday activity patterns [22] (i.e. recurrent time and date specific patterns that correlate with trader behaviour and bias but not to new market information). During step (ii), we briefly evaluate the intraday activity pattern of the studied assets in the evaluated period and choose the trading day from 00:00 to 23:59 as the unit of analysis. This choice is based on the observation that a W-shape intraday-activity curve can be observed starting from 00:00 for BTC, as shown in Fig. 4.

Moreover, Fig. 4 illustrates the existence of trade counts peaks at rounded-hour timestamps. We hypothesize that this pattern is caused by automated batched trade orders.

III. Methods

A. Validation of market sentiment proxy

In this section, we validate our chosen proxy as an explanatory variable for cryptocurrency price changes using transfer entropy as a measure of information flow.

Setup.

IDTxl library was employed to find out which keywords have a significant predictive power on each of the 4 cryptos studied. We assume here that a keyword K provides statistically significant information on an asset C if the estimated transfer entropy $TE_{K\to C}$ passes the surrogate statistical test implemented by *IDTxl* at the 5% significance level.

Bivariate transfer entropy.

IDTxl provides an easy-to-implement tool for estimating the transfer entropy between pairs of source and target processes (X, Y) through the inference of a weighted directed network. Each edge (x, y) is generated if the estimated bivariate transfer entropy from x to y is statistically significant. These edges are weighted by the information-transfer delay (as the lag of the source x with maximum information transfer into the target y) and the estimated transfer entropy.

We inferred a graph by settings all processes (keywords and asset's log-returns) as sources and assets as targets. The results can be found in the results IV-C section.

Multivariate transfer entropy.

Limiting ourselves to studying the predictive power of a single keyword on an asset's variations seems a somewhat restrictive approach. First of all, it's very likely that the combination of several keywords, such as 'btc' + 'China', has a much stronger impact on Bitcoin's log-returns than each keyword taken individually. Furthermore, there may be an information flow between the keywords themselves, biasing the estimates obtained via bivariate transfer entropy. In a more formal context, bivariate analysis may lead to :

- false positives by inferring spurious interactions from one source x to the target, because of the information flow from other sources influencing x,
- false negatives by missing synergistic interactions between multiple relevant sources and the target. These multiple sources may collectively convey more information to the target than analyzing the contributions of individual sources separately.

Because of these limitations, we decided to investigate multivariate transfer entropy to gain a more realistic overview of the interactions between selected keywords and asset's log-returns. We inferred a network where each node can be both a source and a target, enabling us to model the complex dynamics between our chosen proxies for market sentiment and the crypto market.

B. Evolution of lead-lag correlation

Setup.

To compare log-returns, we calculate the lead-time for each pair of assets (X, Y), e.g the. lag for which cross-correlation is the largest between X and Y. Note that a max-lag of 60 seconds was considered. The *Lead-lag* library was employed to infer pairwise Lead-Lag Ratios for each trading day (from 00:00 to 23:59 UTC time) between log-returns of all considered cryptocurrencies and the seven Google Trends keywords yielding statistically significant causal relations under a binary transfer entropy approach ('trading', 'Musk', 'btc', 'Litecoin', 'Ethereum', 'Cardano' and 'FED'). Furthemore, we use pyRMT [23] implementation of RMT correlation matrix cleaning to regularize the correlation matrix of the lead-lag network adjacency matrix, in order to perform clustering. We excluded the trades of Cardano (ADA) on Bitstamp from our analysis. This decision was based on ADA being introduced to Bitstamp relatively late in the studied period, and having either

zero or negligible trade volume for the majority of trading days.

Lead-lag network clustering and dimensionality reduction.

To compare log-returns, we calculate the lead-time for each pair of assets (X, Y), e.g the. lag for which cross-correlation is the largest between X and Y. Note that a max-lag of 60 seconds was considered. We compute pairwise Lead-Lag ratios between (i) all crypto assets, (ii) between Google Trends time series and crypto assets. We exclude autocorrelation estimation from this analysis and focus on cross-correlation between time-series, since our main objective is to assess information flow between market sentiment proxies and cryptocurrency exchanges. We obtain 365 different lead-lag networks, e.g. directed graphs, consisting of 14 nodes and 252 candidate edges. Each network corresponds to one trading day of the studied period. We assess the statistical significance of candidate edges at the 5% significance level against a null model consisting of uncorrelated arithmetic brownian motions and apply the Holm-Sidak correction to control for multiple hypothesis testing error.

In a second step, we are interested in the identification of systematic patterns in the information flow via dimensionality reduction of the 365 previously computed lead-lag networks into a smaller set of network classes. Our approach consists of (i) computing the cleaned correlation matrix between lead-lag edges from the adjacency matrices using Random Matrix Theory (RMT) for each lead-lag network and (ii) applying Maximum Likelihood Clustering [24] on the cleaned correlation matrix of all 365 daily networks.

IV. Results

A. In-Depth Analysis of Bitcoin

Comparison of daily and intraday data. Recall that intraday data was collected through one trading exchange (Binance). To investigate whether the intraday data exhibits similar trends to the daily data acquired through the Yahoo Finance API, we computed basic statistics on the closing price and volume for both daily and intraday data, then visualized their variations 5.

As expected, intraday and daily closing prices are similar. However, the traded volume on Binance does not accurately reflect the total traded volume of Bitcoin. This observation raises a question :

> How do different cryptocurrency exchange platforms behave? Is there considerable disparity in traded volumes? Is there a particular exchange platform that serves as a more representative indicator of the total traded volume?

Bitcoin's analysis. The long-term trajectory of Bitcoin's price has been higher-"up and to the right," as they say - since 2019. Despite the long-term rise, Bitcoin has been dogged by periods where it has fallen precipitously.

Fig. 5: BTC closing price

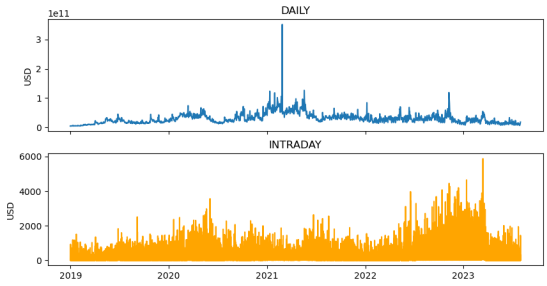


Fig. 6: BTC volume

Fig. 7: Bitcoin's price analysis

2020. The year 2020 begins with a decline in the Bitcoin's price in the midst of the stock market downturns during the initial COVID pandemic wave, with a flash crash on March 12 – from $7'935 to $4'826 in a single day, a decline of more than 39%. Note that the flash crash affected all financial markets, not just Bitcoin.

> "The international stock market crash of March 2020 associated with the first major wave of the COVID-19 pandemic due to a novel coronavirus appears exceptional." [a]
>
> [a] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8450777/

Following this period, the price of bitcoin started to rise, with an exponential looking trend. During this period, called *bull run* in the crypto world, the demand for Bitcoin's increased a lot, suggesting an increase of trust in the asset : prices skyrocketed from 5k on March 12 to 40k on January 8, 2021.

2021. Bitcoin had a strong start in 2021, reaching over

64k in April, fueled by optimism from promises of ongoing liquidity from the Federal Reserve. However, China's warnings in May on cryptocurrency activities led to a significant drop in Bitcoin's value, losing over 50% within months (from 58k in May 8 to 29k in July 19).

> "China in May banned financial institutions and payment companies from providing services related to cryptocurrency transactions." [a]
>
> [a]https://www.reuters.com/world/china/china-central-bank-vows-\ crackdown-cryptocurrency-trading-2021-09-24/

Despite regulatory headwinds in China, Bitcoin recovered, hitting a new all-time high of 68k in November 2021. However, the Federal Reserve's announcement of tapering bond purchases and tightening liquidity on November 3, coupled with concerns about rising inflation, affected various markets, including cryptocurrencies. Bitcoin, along with risky assets, experienced a downturn.

> "The Fed will reduce its current 120 billion monthly bond-buying program by 15 billion a month."[a]
>
> [a]https://www.bankrate.com/banking/federal-reserve/ fomc-meeting-recap-november-2021/)

2022. This downward trend continued into 2022, with Bitcoin fluctuating around $40,000$. As the Federal Reserve aggressively raised interest rates in March 2022, Bitcoin faced further declines, establishing a new trading range around $20,000$ and dropping below $16,000$ due to high-profile issues like the FTX bankruptcy in November 2022, eroding trader confidence.

> "FTX's bankruptcy has left the crypto market weakened, plagued by a crisis of confidence and regulatory tightening."[a]
>
> [a]Translated from www.lesechos.fr/finance-marches/marches-financiers/ ftx-un-crash-massif-et-cinq-consequences-pour-lindustrie-crypto-1983868

2023. In 2023, the price experienced an upswing, surging by over 50 percent until mid-June, driven by a general upward trend in technology stocks. Despite regulatory measures by the Securities and Exchange Commission targeting the cryptocurrency industry, Bitcoin was trading at approximately $26,000 as of mid-June 2023.

Comparison with traditional financial markets. To gain insights into where Bitcoin stands relative to traditional financial markets, we conducted basic analysis on two representative assets: S&P 500, AAPL and reported the main results in Table II.

Asset	mean	std	std/mean * 100
BTC	$23'850	$16'478	69%
S&P 500	$3'690	$624	17%
AAPL	$116	$45	39%

TABLE II: Statistics on closing prices

Bitcoin closing prices are on average much higher across the studied time period (31-12-2018 to 01-08-2023). Additionally, the standard deviation observed for Bitcoin is significantly larger (approximately 69% of the mean value) compared to the selected indicators of traditional financial markets (17% for the S&P 500, 39% for Apple). This discrepancy underscores higher volatility and elevated risk in Bitcoin.

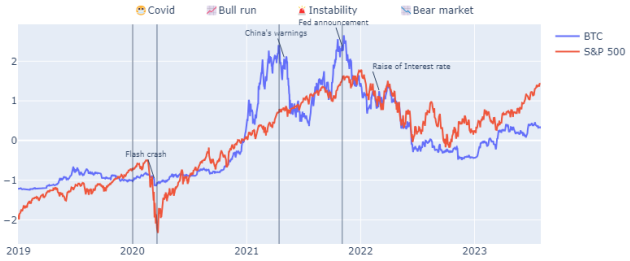


Fig. 8: BTC - S&P 500

It is evident that Bitcoin's trajectory is closely linked to broader financial market movements, highlighting furthermore the crypto market's sensitivity to major economic events. However, from mid-2020 to the end of 2021, Bitcoin displayed more peaks and troughs compared to the relatively stable and linear upward trend observed in the S&P 500 index. This divergence can be attributed to heightened market attention and increased scrutiny of cryptocurrency by society, as well as news specifically impacting the crypto market, such as China's involvement.

Conclusion. Following the results we just obtained, we have decided to focus on the period from 01-01-2021 to 31-12-2021. Indeed, this year was marked by many significant events which had a major impact on Bitcoin's price. Our period of interest is characterized by heightened tension, high volatility and numerous shifts in market sentiment. In fact, at the beginning of the year 2021, confidence in BTC was at its peak, leading to exceptional growth, but this was followed by a period of intense instability following China's announcement.

It seems particularly relevant to study this period further, as it appears to be one of the stages when market sentiment towards cryptocurrencies was at its strongest. Consequently, having a stronger information flow makes it more likely for us to evaluate the impact of market sentiment on the cryptocurrency market.

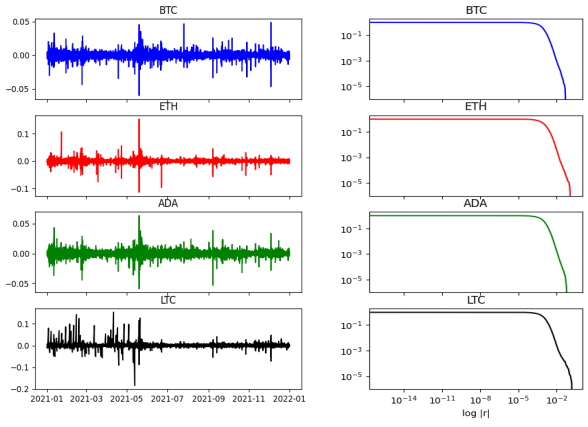
B. Stylized facts on Price Movements

As part of the data exploration and quality assessment steps of our data processing pipeline, we verify known stylized facts of the finance literature as a preliminary analysis step. We check for (i) heavy-tailedness of log-returns of all cryptocurrencies (while aggregating data coming from both exchanges), (ii) absence of autocorrelation of volatility and long-memory

of volatility, (iii) slowly-decreasing response function of log-returns with respect to trade initiation for long lags.

Log-returns

The log-returns of the 4 studied cryptos have similarities. Notably, a cluster of high volatility is evident around May 2021, possibly due to the uncertainty in the market triggered by the announcements from China leading to a significant drop in Bitcoin's value. Additionally, the hypothesis of a power-law distribution seems plausible for all assets after visual examination of the complementary ecdf in log-log scale. To validate these observations, statistical tests were conducted with the *powerlaw* library [5].



(a) Log-returns (b) Complementary ecdf

Fig. 9: Logreturns time-series and ECDF for considered assets

	BTC	LTC	ADA	ETH
alpha	4.38	3.50	3.95	3.59
xmin	0.0007	0.0004	0.0008	0.0006

TABLE III: Estimated parameters of powerlaw.

Asset	Results	Lognormal	Exponential
BTC	z-score	-0.219	1.729
	p-value	0.826	0.08
	Powerlaw favored ?	False	False
LTC	z-score	-3.776	0.965
	p-value	**0.0002**	0.335
	Powerlaw favored ?	False	False
ADA	z-score	-1.240	1.308
	p-value	0.215	0.191
	Powerlaw favored ?	False	False
ETH	z-score	0.436	4.308
	p-value	0.663	**0.00002**
	Powerlaw favored ?	False	True

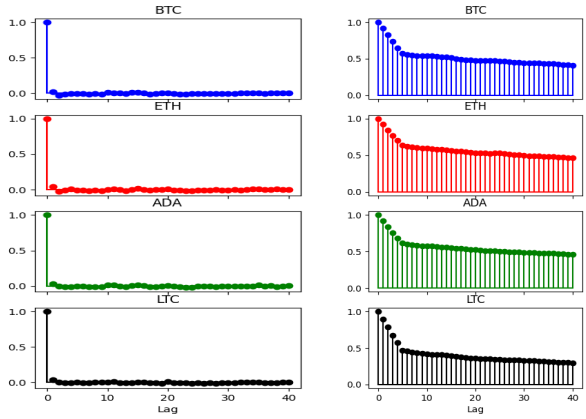
TABLE IV: Results from statistical tests.

Table III shows the estimated coefficients of the power-law distribution for each asset. The alpha exponent remains relatively homogeneous across crypto, slightly surpassing the average value of $\alpha \approx 3$ observed in classical financial data.

[5]https://pypi.org/project/powerlaw/

Table IV displays the results of the statistical tests, which aim to determine the preferred distribution for each asset. It is noteworthy that most tests reject the hypothesis that log-returns follow a power-law distribution. It is in fact significantly favored over exponential only for ETH. For LTC, the results suggest a statistically significant preference for lognormal distribution, while for BTC and ADA, none of the tests produced a p-value lower than 0.05, indicating that it is impossible to distinguish between a power-law, lognormal or exponential distribution.

Autocorrelation and volatility. We decided to fix the window size to 5 for the computation of volatility, our choice being driven by our specific interest in capturing the short-term fluctuations in log-returns.



(a) ACF of log-returns (b) ACF of volatility

As expected, we don't observe linear autocorrelation in logreturns at different lags for all assets.
Furthermore, the ACF of volatility is decreasing slowly, highlighting long-memory property of volatility. In other terms, the volatility tends to exhibit an autocorrelated and persistent behavior over time, displaying patterns of clustering over time.

Response Function.

We analyze the predictive performance of signs of trade initiators on market order book midprices via response functions. The response functions for Binance traded BTC, ETH, ADA and LTC are depicted in Figure 11. We abridge the discussion to the response function on market order of the 1st January 2021 and for lags of 100 traded as illustrative examples of the general functional form of the response function, even if we observe variations accross different days. We can observe that ETH, LTC and ADA response functions present a peak in 300-500 lagged trade counts, meaning that trade initiation sign impact on trades has long memory. In the other hand, BTC response function presents a peak in 10 trades, and achieves values up to 10 times lower than the response function of other coins, possibly to the higher trade volume of Bitcoin when compared to other cryptocurrencies.

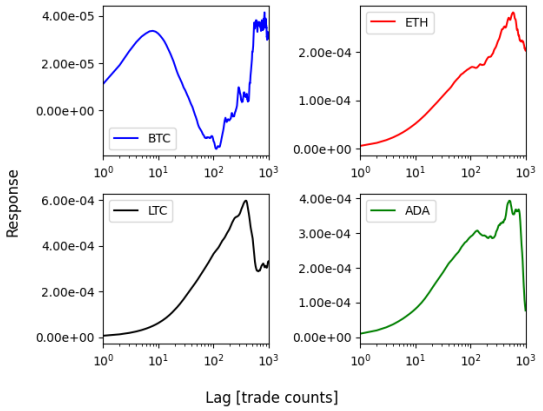


Fig. 11: Response function for order book mid-prices with respect to trade initiator sign.

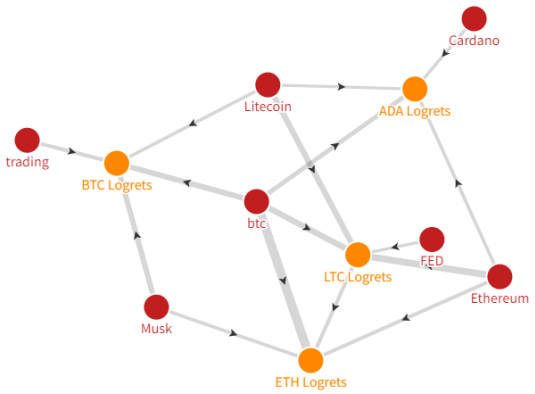


Fig. 13: bTE inferred graph - lags in [1,2,3,4]

C. *Transfer Entropy from keywords to crypto log-returns.*

Bivariate Transfer Entropy.

We first estimate the bivariate transfer entropy between each pair of processes $\{(X,Y)|X \in \{\text{Keywords} \cup \text{Logrets}\} \setminus \{Y\}, Y \in \{\text{Logrets}\}\}$.

Note that we have restricted the maximum lag, e.g. the number of past values considered for each process, to 4, and set the minimum lag to 1. Given that we are working with hourly data, we first wanted to understand the 'short-term' dynamics between our market sentiment proxies and the variations in assets by studying the information-delay (in the 1-4 hour range) between pairs of processes that show a statistically significant interaction.

Figure 12 illustrates these concepts. In this example, the minimum and maximum lag for the source processes in X are l_3 and l_1 respectively, while the maximum lag for the target Y is l_2. The minimum target lag is always set to 1, to ensure so-called self-prediction optimality [15].

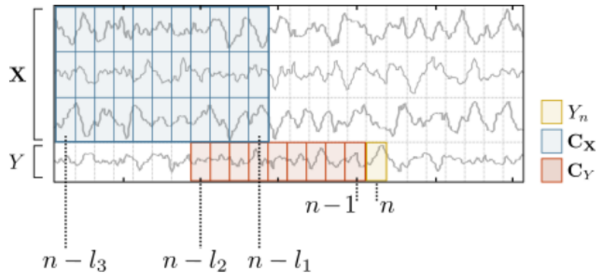


Fig. 12: Maximum/minimum lag.

We obtained the network depicted in 13, where nodes represent time-series and arrows represent interactions between processes. The width of an edge is proportional to the value of the transfer entropy from source to target.

Out of the 16 keywords examined, only 7 are retained, from which many keywords demonstrate a predictive capacity on several assets simultaneously (such as *'btc'*, *'Ethereum'*, *'Litecoin'*). This indicates that despite the small number of keywords selected, the latter are closely linked to the cryptocurrency market, exerting a significant influence on several assets simultaneously and highlighting the complex dynamics of this interconnected network.

It is interesting to note that there is information transfer between each keyword directly related to an asset : for instance, estimated $TE_{\text{btc}\to\text{BTC\_Logrets}}, TE_{\text{Ethereum}\to\text{ETH\_Logrets}}$ are statistically significant. The inferred graph also highlights the interactions between assets within the cryptocurrency market itself, with LTC's log-returns directly influencing ETH's log-returns. Furthermore, many asset keywords (e.g., *'Litecoin'*, *'Ethereum'*, *'btc'*) have predictive power over the variations of other assets, with for example significant estimated $TE_{\text{Litecoin}\to\text{ADA\_Logrets}}$. More details about the inferred results are reported in appendix VI-B.

Figure 14 summarizes the information delay between each pair (source, target) in the inferred graph. Recall that the information-transfer delay [13] between 2 processes (X,Y) is estimated as the time lag between X's past value and the current value of Y which provides the maximum TE estimate. Note that observed lags were restricted to the interval $[1,4]$. Lag 4 predominates the results, representing 40% of lags with highest TE estimate. Unexpectedly, lag 1 is optimal for only 3 links. This observation prompted us to infer a new graph, with a higher maximum lag, in order to find out whether keywords' influence on the crypto market is more based on long-term dynamics.

Our second model is based on slightly different settings than the previous one, with the aim of focusing on higher lags. Since the inference process is computationally intensive, we have chosen to restrict the test to a limited number of lags, while examining lags of up to 22. To achieve this, *IDTxl* provides a parameter, τ, which allows us to include only a subset of past observations in the source process, selecting each τ^{th} observation until the maximum lag is reached. In our

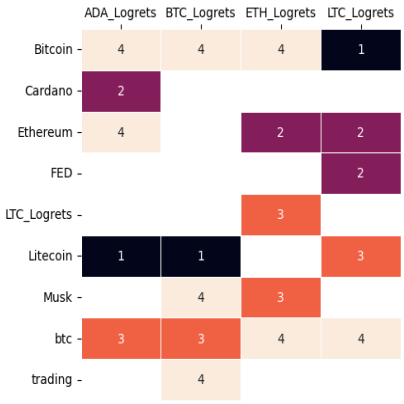


Fig. 14: Information delay - lags in [1,2,3,4]

setting, τ was set to 6 and we looked at lags in $\{4, 10, 16, 22\}$. Obtained results are reported below in Fig. 15.

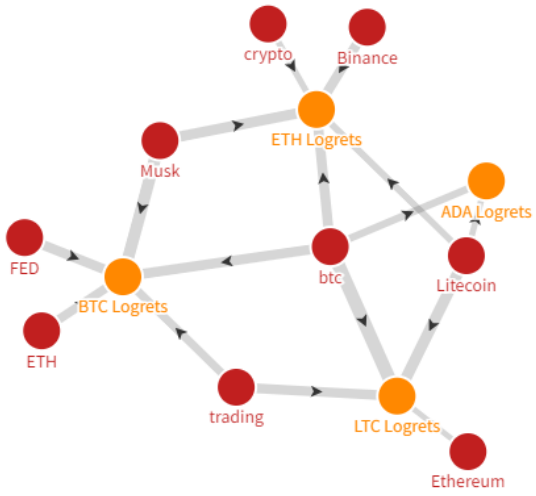


Fig. 15: bTE inferred graph - lags in [4,10,16,22]

The first significant difference with the previous model is that for ETH and ADA, no keyword directly associated with these cryptos ('ETH', 'Ethereum', 'ADA', 'Cardano') has a direct influence on them. However, the estimated transfer entropy from other crypto-related keywords, such as 'btc' or 'Litecoin', to these assets is statistically significant. The results are somewhat less "coherent" with the general idea of associating a keyword directly linked to a crypto with its variations. However, since we're dealing with higher lags, this may be an underlying effect of the crypto market itself.

Some keywords, such as 'Binance', now appear in this second graph. However, it's important to note that the majority of the keywords selected here are the same as before, as well as a majority of the links - for instance 'btc' influencing the 4 assets' log-returns. This observation accentuates our hypothesis that some keywords are strong indicators of market sentiment towards cryptocurrencies, which also appear to be consistent across lags.

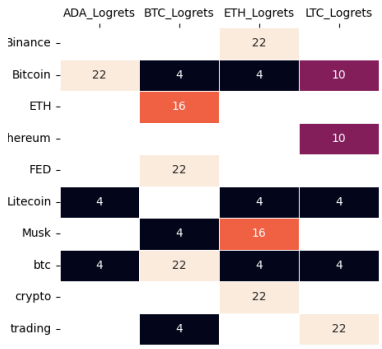


Fig. 16: Information delay - lags in [4,10,16,22]

The time lag that yields the maximum Transfer Entropy (TE) estimate is 4 for 50% of the statistically significant pairs of processes. However, a substantial number of keywords still exert their maximal influence at lag 22 and some keywords present in the first model with an information delay of 4 have now an information delay of 22.

Multivariate Transfer Entropy.

As introduced in section III-A, limiting ourselves to study the predictive power of a single keyword on asset's variations seems a somewhat restrictive approach. That's why, to complete this analysis, we decided to investigate multivariate transfer entropy to gain a more realistic overview of the interactions between selected keywords and asset's variations.

The objective here is to infer a network from the log-returns of all considered cryptocurrencies and Google Trends time-series. To do so, instead of just evaluating the bivariate transfer entropy $TE_{X\to Y}$ between pairs of processes $\{(X,Y)|X \in \{\text{Keywords} \cup \text{Logrets}\} \setminus \{Y\}, Y \in \{\text{Logrets}\}\}$, we consider each time-series as both a source and a target. This approach enables us to get a better understanding of the intricacy and complexity of the information flow between the selected keywords and cryptocurrencies.

For the inference of the network and its node dynamics, we used the algorithm implemented in the *IDTxl* library, based on the work of Lizier and Faes [25]. This algorithm infers, for each node treated as a target, all relevant sources by iteratively including variables that maximise a conditional mutual information criterion. This iterative conditioning is designed to both eliminate redundancies and capture synergistic interactions in building each parent set, thus addressing the issues of bivariate analysis. The obtained results are summarized in Figs. 17 and 16.

(i) A complex network

While 7 keywords (lags 1 to 4) then 9 (lags 4, 10, 16, 22) were statistically significant in predicting the variations of our cryptos under study in the bTE inferred graphs, the network resulting from multivariate entropy allows us to model a

far more complex flow of information between the set of keywords selected at the beginning of the project. Instead of limiting ourselves to analyzing the impact of each keyword on each crypto individually with bivariate transfer entropy, we are now able to understand in a more in-depth (and causally correct) way how the popularity of certain keywords on the Internet influences the cryptocurrency market, while at the same time highlighting the interactions between our proxies.

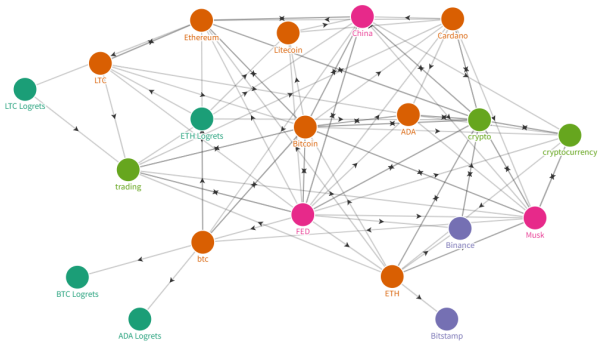


Fig. 17: mTE inferred graph - lags in [1,2,3,4]

(ii) Bivariate entropy yields causally inaccurate results. A significant number of keywords that passed the statistical test for bivariate transfer entropy (bTE) lose their statistical significance when multivariate entropy (mTE) is considered. Indeed, only the keywords 'btc' and 'Ethereum' show a direct and statistically significant impact on all four cryptocurrencies studied. Estimating the transfer entropy between each pair of processes independently via the bTE method may have overestimated the predictive power of most of the selected keywords, the latter merely reflecting the impact of the retained keywords via mTE method.

Consider the case of ETH's Logrets for a better understanding of this phenomenon. When applying the bTE method with lags ranging from 1 to 4, 'Bitcoin', 'Ethereum', 'Musk' and 'btc' keywords are selected, while only 'btc' is retained for the mTE method. To gain a deeper insight into the relationships between these keywords in the network inferred by mTE, we examine the subgraph $\mathcal{G}_{\text{btc}}$, which only includes parents and children of node 'btc'.

'btc' turns out to be a confounder for the keywords 'Bitcoin' and 'Ethereum' with respect to ETH Logrets, i.e. it simultaneously causes 'Bitcoin'/'Ethereum' and ETH Logrets through the paths 'Bitcoin' $\leftarrow$ 'btc' $\rightarrow$ ETH Logrets and 'Ethereum' $\leftarrow$ 'btc' $\rightarrow$ ETH Logrets. With the bTE method, we have falsely inferred a direct causal relation from 'Bitcoin' (and 'Ethereum') to ETH Logrets, when in fact these keywords only reflect the impact of the keyword 'btc' on the variations of this asset. This problem stems from the famous adage: "Association is not causation". The keywords 'Bitcoin' and 'Ethereum' are strongly associated with 'btc', the only one with a direct causal effect on ETH's Logrets, but do not directly cause Ethereum's variations.

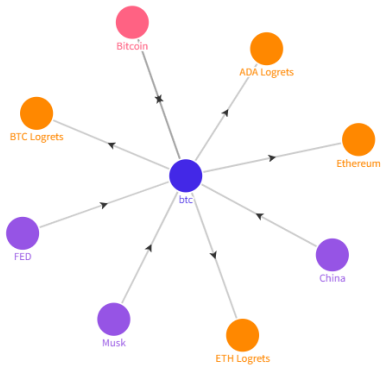


Fig. 18: Subgraph $\mathcal{G}_{\text{btc}}$

Meanwhile, the keyword 'Musk' is linked to ETH Logrets via the path 'Musk' $\rightarrow$ 'btc' $\rightarrow$ ETH Logrets. The configuration differs here: 'Musk' causes ETH's Logrets, but only through its influence on 'btc'. Using conditional mutual information between 'Musk' and ETH Logrets by conditioning on 'btc' has rendered 'Musk' statistically insignificant, indicating that his predominant influence was mainly on 'btc' and not on cryptocurrency variations.

The mTE inferred graph is therefore far more reliable than the other models, which could have led to causally false conclusions. This basic network shows the complexity of determining causal relationships and quantifying the flow of information between a specific keyword and crypto log-returns, given that they are part of a vast, dynamic, complex and interconnected network.

However, the fact that 'btc' is the only keyword with a direct impact on 3 of the 4 studied assets highlights how it is a strong indicator of market sentiment, with significant predictive power on the crypto market.

(iii) Significant keywords for understanding Market Sentiment Dynamics

While the log-returns of each cryptocurrency have only one parent in the mTE inferred graph and no descendants, certain keywords exhibit significant importance in the network based on well-known centrality indicators such as in-degree and out-degree.

In-degree	Out-degree
Bitcoin (12)	FED (14)
crypto (11)	Bitcoin (12)
Musk (9)	crypto (12)
Ethereum (8)	China (7)

TABLE V: Top 4 nodes with highest in-degree and out-degree.

Keywords like 'Bitcoin' and 'crypto' demonstrate substantial in and out degrees, ranking as the first two nodes with the highest number of ingoing edges (12 and 11, respectively)

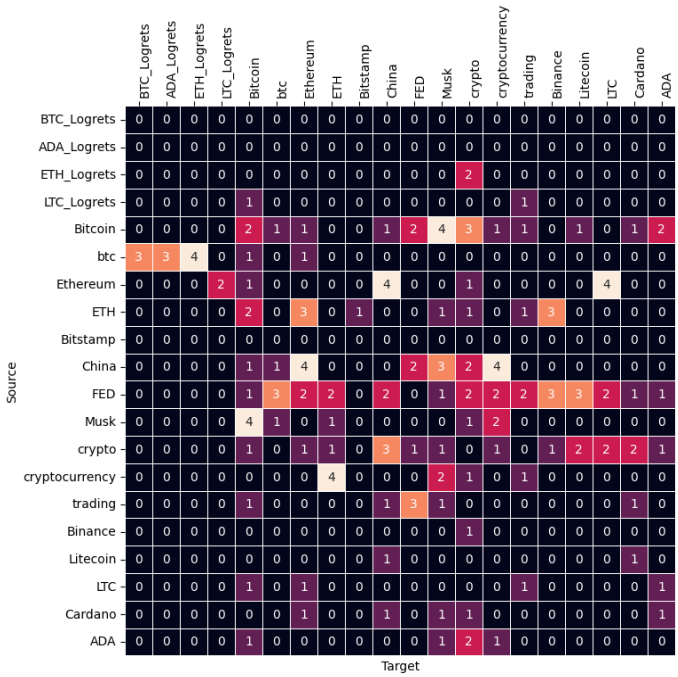


Fig. 19: mTE information delay

and appearing in the top four nodes with the highest number of outgoing edges. These keywords appear to hold a central position, acting as both the cause and consequence of more than half of the processes present in the network (20).

Other keywords, such as 'China', 'FED' and 'Musk', previously identified as impactful when studying the global variations of Bitcoin in part IV-A, also occupy dominant positions in the mTE network, according to our centrality indicators. Even if these keywords have no direct impact on cryptocurrency movements, they are essential in understanding market sentiment towards the crypto market. It is therefore essential not to underestimate the importance of these keywords.

We could also have used other centrality indicators and analyzed the resulting graph in more detail, but this is beyond the scope of the project.

D. Lead-lag networks

(i) Lead-lag relations may span several seconds.

Representative lead-lag networks of the clustering classes found by our model are depicted in Fig. 20, while the complete set of 365 graphs are available in our code repository. In these examples, we consider trade-by-trade log-returns for BTC, ETH, LTC in both Binance and Bitstamp exchanges and ADA log-returns only in the Binance exchange, since Cardano was not traded in Bitstamp during most of the considered time period. We consider only the seven statistical significant keywords according to binary transfer entropy measurements ('trading', 'Musk', 'btc', 'Litecoin', 'Ethereum', 'Cardano' and 'FED').

The first example in Fig. 20 corresponds to the lead-lag graph of log-returns and market sentiment on the 1st January 2021. During this particular day, we observe a significant lag of returns of LTC when traded in Binance compared to the same coin traded in Bitstamp, or other cryptocurrencies of the basket. The second example correspond to the lead-lag graph of 20 July 2021. In this second example, many lead-lag relations are statistically significant, specially links in which ADA and LTC coins log-returns are the follower time series. In general, we observe that BTC and ETH log-returns between exchanges do not consistently exhibit a lead-lag relation, while LTC attains the largest and most volatile LLR whe considering all trading days, with a mean lag-time of 2.2 seconds.

(ii) Bitstamp traded assets present significantly more lead-lag relations than Binance traded assets.

In order to assess the main characteristics of our networks, we use in-degree (and out-degree) centrality indicators as estimators of the tendency of an asset to lead (follow) changes in time-series of other assets. As summarized in Table VII, Bitstamp is, overall, the exchange containing assets most often in a "leader" and "follower" positions in the network. This could possibly indicate that Bitstamp cryptocurrency prices responds more to changes in other quote prices and market sentiment time-series. Moreover, we observe that LTC is the most frequent coin being in the follower position. We summarize further results, such as the persistence of links, in Table VI.

(iii) Correlation matrix cleaning regularizes maximum likelihood clustering. Results of the Maximum Clustering approach used to asses the existence of "information flow states" in our time-series of lead-lag networks are depicted in Fig. 21 and 22 for empirical correlation and RMT cleaned empirical correlation respectively. In Fig. 21, we observe that maximum likelihood clustering produces 12 different clusters. If we implement eigenvalue cleaning on the graphs' adjacency correlation matrix with RMT, we obtain only 3 states. We are not able to observe the effect of the main 2021 events on the Bitcoin market as describe in section IV-A with this technique and hypothesize that it is possible that the classes observed correspond to exchange- or coin-specific market states for which interpretation is not straightforward.

Summary of main results. When considering trade-level data aggregated by day and separated by exchange analyzed with lead-lag approaches, we observed that causal relations are of the order of milliseconds to seconds, with the longest statistical significant relation spanning slightly above 2 seconds. We interpret this fact as an indicator that cryptocurrency trade-by-trade returns are highly volatile and encode new information very rapidly. We further observed that cryptocurrencies traded on Bitstamp participate in more lead-lag relations than Binance traded assets. A possible explanation for this observations is the difference in traded volume between both exchanges: since Bitstamp has significant less trades than Binance, information takes more time to flow and leading-lag relations that are otherwise too fast to measure or lost in noise become observable.

Finally, we observed that the unsupervised clustering of daily lead-lag networks with a maximum likelihood approach is a challenge. While Maximum Likelihood clustering allows

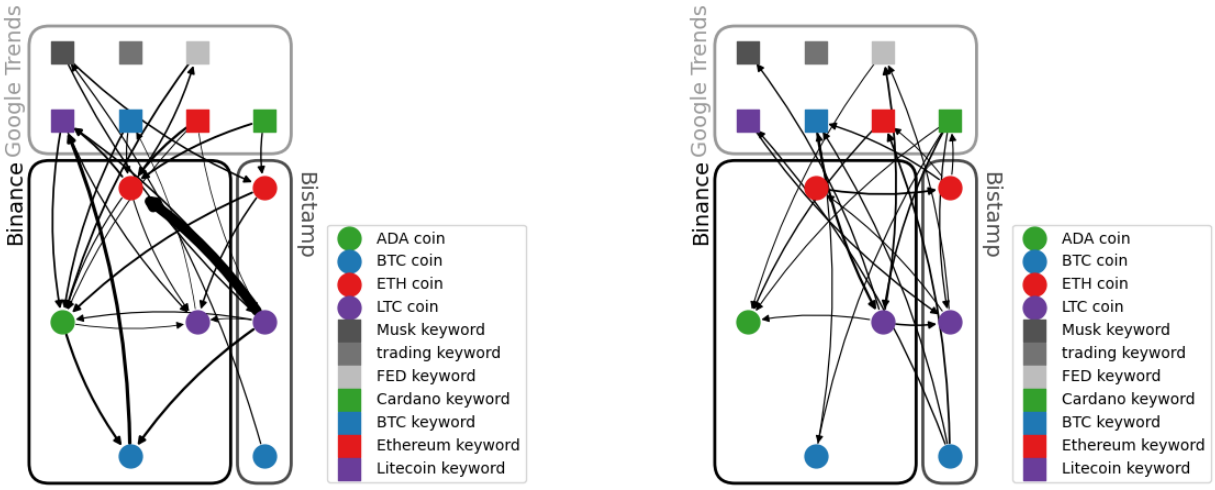


Fig. 20: Lead-lag networks for 1st January 2021 (left) and 20 July 2021 (right). Links are statistically significant with 5% confidence under Holm-Sidak correction and arrow thickness is proportional to LLR.

us to separate trading days in 12 "market states", it does not provide a clear interpretation of main characteristics of the clusters considered, since variations between days in a cluster can be of the order of magnitude as the variations between clusters. Correlation matrix cleaning with RMT reduces the number of clusters to 3 and is helpful as a regularizer when the objective of clustering is dimensionality reduction, but does not improves interpretability of the clustering method.

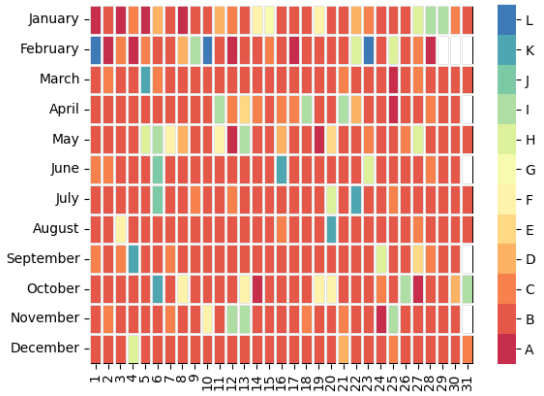


Fig. 21: Lead-Lag Network clusters without RMT cleaning

V. Discussion

In this study, we conducted a comprehensive exploration of information flow dynamics within the cryptocurrency market, with a focus on understanding the interactions between different exchanges, market sentiment proxies, and cryptocurrency assets. Despite the challenges posed by limited access to tick-by-tick trade data, our investigation into stylized facts related to crypto-currency returns and the intricate patterns of information flow yielded valuable insights.

Our approach involved the selection of Google Trends keywords as market sentiment proxies and the application of

Fig. 22: Lead-Lag Network clusters with RMT cleaning

both entropy-based methods and lead-lag analysis to unravel the causal relationships and temporal dependencies between these proxies, cryptocurrency exchanges, and asset prices.

The entropy-based analyses, specifically bivariate and multivariate transfer entropy, revealed a complex network of interactions. We observed the influence of selected keywords on cryptocurrency variations, identifying 'btc' as a robust indicator with a direct impact on multiple assets. The introduction of multivariate transfer entropy highlighted the nuanced nature of causality, leading to refined and more accurate insights. Notably, certain keywords that appeared significant in bivariate analyses lost their significance when considering a broader context, emphasizing the importance of a multivariate perspective.

Our lead-lag analysis, scrutinizing the relations between exchanges and selected assets, uncovered rapid relationships spanning milliseconds to seconds. Bitstamp emerged as a key player, participating in more lead-lag relations than Binance, potentially due to differences in trading volumes.

	BTC (Binance → Bitstamp)	ETH (Bitstamp → Binance)	LTC (Binance → Bitstamp)
(Annual.) Mean LLR	1.0751	1.1997	1.6134
(Annual.) Std. Dev. LLR	0.4076	0.223	0.6800
Max. LLR	3.7972	7.6970	5.785
Median LLR	1.0511	1.1662	1.4908
(Annual.) Mean Lag time [s]	0.36	0.43	2.21
(Annual.) Std. Dev. Lag time [s]	4.48	2.03	5.80
Max. Lag time [s]	43.4	52.8	46.0

TABLE VI: Lead Lag ratio and Lead time statistics of log-returns for a fixed cryptocurrency between two different exchanges, computed in the edge direction attaining the largest LLR.

Description	Statistic
Maximum In-Degree node Mean ± Std. In-Degree	Bitstamp LTC, 4.36 ± 2.46
Maximum Out-Degree node Mean ± Std. Out-Degree	Bitstamp ETH, 2.44 ± 1.68
Most often leader exchange Mean ± Std. In-Degree	Bitstamp, 2.95 ± 1.24
Most often follower exchange Mean ± Std. Out-Degree	Bitstamp, 2.05 ± 1.52
Mean link persistence	4.2 consecutive days
Max. link persistence	69 consecutive days Bin. LTC → Bits. LTC

TABLE VII: Lead-Lag ratio degree statistics.

Despite the inherent challenges in interpreting large-scale events in relation to information flow networks, our study provides valuable insights into the temporal dynamics of the cryptocurrency market. We notice that RMT is a powerful technique to regularize correlation matrices, which allowed decreasing our unsupervised cluster counts from 12 to 3, but explainability of unsupervised maximum likelihood clustering remains a challenge.

Looking ahead, this work opens avenues for future research, including the exploration of additional signals such as price volatility or trade volume, expanding the number of assets and exchanges considered, and enhancing the interpretability of clustering approaches.

In conclusion, our study contributes to the evolving understanding of information flow within the cryptocurrency market, emphasizing the need for nuanced analyses that consider the interplay of various factors. As the crypto landscape continues to evolve, further investigations into these dynamics will be essential for gaining deeper insights into market behavior and facilitating informed decision-making.

ACKNOWLEDGEMENTS

We extend our gratitude to Pr. Damien Challet and Federico Lanfranchi for their valuable assistance and insightful contributions throughout the writing of this work.

VI. APPENDIX

A. IDTxl's greedy algorithm for mTE and bTE estimation

1) Initialise Z as an empty set.
2) Find all relevant variables in the target's own past.
   - For each lag $\tau \in \{1, \text{max\_lag\_target}\}$, estimate the information contribution from $Y_{t-\tau}$ to Y_t.
   - Find the candidate Y'_t with maximum transfer entropy and perform a significance test.
   - If not significant or no candidates, go to step 3. Otherwise, append Y'_t to Z and repeat step 2 with a new candidate.
3) Find all relevant variables in the source's own past.
   - For each lag $\tau \in \{1, \text{max\_lag\_source}\}$, estimate the information contribution from $X_{t-\tau}$ to Y_t
   - Find the candidate with maximum transfer entropy and perform a significance test.
   - If not significant or no candidates, go to step 3. Otherwise, append X'_t to Z and repeat step 3 with a new candidate.
4) Test and remove redundant variables in Z by computing the conditional mutual information between every selected variable in Z and the Y_t, conditional on all the remaining variables in Z [26]
5) Conduct statistical tests on the final variable set Z:
   - Test the combined information transfer from all source variables to the target.
   - If the collective test yields significance, proceed to individual tests on each variable $z \in Z$ to determine each variable's final information contribution and associated p-value.

B. Inference results.

REFERENCES

[1] J. V. Critien, A. Gatt, and J. Ellul, "Bitcoin price change and trend prediction through twitter sentiment and data volume," *Financial Innovation*, vol. 8, no. 1, p. 45, May 2022. [Online]. Available: https://doi.org/10.1186/s40854-022-00352-7

[2] S. McNally, J. Roche, and S. Caton, "Predicting the price of bitcoin using machine learning," in *2018 26th Euromicro International Conference on Parallel, Distributed and Network-based Processing (PDP)*, 2018, pp. 339–343.

Target	Source	Info delay	Estimated bTE	p-value
BTC Logrets	Bitcoin	4	0.026238	0.002
	btc	3	0.028599	0.002
	Musk	4	0.026932	0.002
	trading	4	0.019685	0.006
	Litecoin	1	0.017150	0.010
ADA Logrets	Bitcoin	4	0.020975	0.002
	btc	3	0.023396	0.002
	Ethereum	4	0.016977	0.022
	Litecoin	1	0.015580	0.016
	Cardano	2	0.016350	0.026
ETH Logrets	LTC Logrets	3	0.019126	0.012
	Bitcoin	4	0.025819	0.002
	btc	4	0.046540	0.002
	Ethereum	2	0.019592	0.004
	Musk	3	0.017782	0.016
LTC Logrets	Bitcoin	1	0.029083	0.002
	btc	4	0.030690	0.002
	Ethereum	2	0.035463	0.002
	FED	2	0.015778	0.026
	Litecoin	3	0.030237	0.002

TABLE VIII: bTE - lags in [1,2,3,4].

Target	Source	Info delay	Estimated bTE	p-value
BTC Logrets	Bitcoin	4	0.026025	0.002
	btc	22	0.022892	0.002
	ETH	16	0.020502	0.008
	FED	22	0.022693	0.002
	Musk	4	0.027561	0.002
	trading	4	0.018701	0.010
ADA Logrets	Bitcoin	22	0.014723	0.042
	btc	4	0.016985	0.014
	Litecoin	4	0.014048	0.028
ETH Logrets	Bitcoin	4	0.018514	0.004
	btc	4	0.023103	0.002
	Musk	16	0.024241	0.002
	crypto	22	0.014518	0.028
	Binance	22	0.022398	0.004
	Binance	4	0.015516	0.028
LTC Logrets	Bitcoin	10	0.025258	0.002
	btc	4	0.030508	0.002
	Ethereum	10	0.012931	0.002
	Ethereum	16	0.013359	0.002
	trading	22	0.022304	0.004
	Litecoin	4	0.025917	0.002

TABLE IX: bTE - lags in [4,10,16,22].

[3] Y. Li and W. Dai, "Bitcoin price forecasting method based on cnn-lstm hybrid neural network model," *The Journal of Engineering*, vol. 2020, no. 13, pp. 344–347, 2020. [Online]. Available: https://ietresearch.onlinelibrary.wiley.com/doi/abs/10.1049/joe.2019.1203

[4] Wikipedia contributors, "Binance — Wikipedia, the free encyclopedia," 2023, [Online; accessed 3-December-2023]. [Online]. Available: https://en.wikipedia.org/w/index.php?title=Binance&oldid=1187809678

[5] ——, "Bitstamp — Wikipedia, the free encyclopedia," 2023, [Online; accessed 3-December-2023]. [Online]. Available: https://en.wikipedia.org/w/index.php?title=Bitstamp&oldid=1179296812

[6] Y. Hu, Y. G. Hou, and L. Oxley, "What role do futures markets play in bitcoin pricing? causality, cointegration and price discovery from a time-varying perspective?" *International Review of Financial Analysis*, vol. 72, p. 101569, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1057521920302131

[7] P. Augustin, A. Rubtsov, and D. Shin, "The impact of derivatives on spot markets: Evidence from the introduction of bitcoin futures contracts," Frankfurt a. M., LawFin Working Paper 41, 2022. [Online]. Available: http://hdl.handle.net/10419/262362

[8] N. Woloszko, "Tracking activity in real time with google trends," no. 1634, 2020. [Online]. Available: https://www.oecd-ilibrary.org/content/paper/6b9c7518-en

[9] T. Preis, H. S. Moat, and H. E. Stanley, "Quantifying trading behavior in financial markets using google trends," *Scientific Reports*, vol. 3, no. 1, p. 1684, Apr 2013. [Online]. Available: https://doi.org/10.1038/srep01684

[10] D. Challet and A. B. H. Ayed, "Predicting financial markets with google trends and not so random keywords," 2014.

[11] C. W. J. Granger, "Investigating causal relations by econometric models and cross-spectral methods," *Econometrica*, vol. 37, no. 3, pp. 424–438, 1969. [Online]. Available: http://www.jstor.org/stable/1912791

[12] R. F. Engle and C. W. J. Granger, "Co-integration and error correction: Representation, estimation, and testing," *Econometrica*, vol. 55, no. 2, pp. 251–276, 1987. [Online]. Available: http://www.jstor.org/stable/1913236

[13] T. Schreiber, "Measuring information transfer," *Physical Review Letters*, vol. 85, no. 2, p. 461–464, Jul. 2000. [Online]. Available: http://dx.doi.org/10.1103/PhysRevLett.85.461

[14] C. Bongiorno and D. Challet, "Statistical inference of lead-lag at various timescales between asynchronous time series from p-values of transfer entropy," Jun 2022. [Online]. Available: https://arxiv.org/abs/2206.10173

[15] M. Wibral, N. Pampu, V. Priesemann, F. Siebenhühner, H. Seiwert, M. Lindner, J. T. Lizier, and R. Vicente, "Measuring information-transfer delays," *PLoS One*, vol. 8, no. 2, p. e55809, Feb 2013. [Online]. Available: https://doi.org/10.1371/journal.pone.0055809

[16] B. Tóth and J. Kertész, "The epps effect revisited," *Quantitative Finance*, vol. 9, no. 7, p. 793–802, Oct. 2009. [Online]. Available: http://dx.doi.org/10.1080/14697680802595668

[17] T. Hayashi and N. Yoshida, "On covariance estimation of non-synchronously observed diffusion processes," *Bernoulli*, vol. 11, no. 2, pp. 359–379, 2005. [Online]. Available: http://www.jstor.org/stable/3318933

[18] N. Huth and F. Abergel, "High frequency lead/lag relationships — empirical facts," *Journal of Empirical Finance*, vol. 26, pp. 41–58, 2014. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0927539814000048

[19] B. Anderson, "A tick-by-tick level measurement of the lead-lag duration between cryptocurrencies: The case of bitcoin versus cardano," *Investment Management and Financial Innovations*, vol. 20, no. 1, p. 174–183, Feb. 2023. [Online]. Available: http://dx.doi.org/10.21511/imfi.20(1).2023.15

[20] H. Schnoering and H. Inzirillo, "Deep fusion of lead-lag graphs: Application to cryptocurrencies," 2022.

[21] G. M. Caporale and A. Plastun, "The day of the week effect in the cryptocurrency market," *Finance Research Letters*, vol. 31, 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1544612318304240

[22] J. P. Broussard and A. L. Nikiforov, "Human bias in algorithmic trading," *SSRN Electronic Journal*, 2013. [Online]. Available: http://dx.doi.org/10.2139/ssrn.2375739

[23] G. Giecold and L. Ouaknin, "pyrmt," https://github.com/GGiecold/pyRMT, 2017.

[24] L. Giada and M. Marsili, "Algorithms of maximum likelihood data clustering with applications," *Physica A: Statistical Mechanics and its Applications*, vol. 315, no. 3–4, p. 650–664, Dec. 2002. [Online]. Available: http://dx.doi.org/10.1016/S0378-4371(02)00974-3

[25] J. T. Lizier and M. Rubinov, "Multivariate transfer entropy," *Preprint, Technical Report 25/2012, Max Planck Institute for Mathematics in the Sciences*, 2012. [Online]. Available: http://www.mis.mpg.de/preprints/2012/preprint2012_25.pdf

[26] ——, "Multivariate construction of effective computational networks from observational data," *arXiv preprint arXiv:1205.0694*, May 3 2012.