Analysing Performance of Stocks using Causal Discovery

Dhanalaxmi Gaddam[1], Diego Saenz[1], and Rushali Grandhe[1]

[1]Mohamed bin Zayed University of Artificial Intelligence
{firstname.lastname}@mbzuai.ac.ae

Abstract

Stock prices data is considered by many as undecipherable and noisy due to its volatility and multi-variable interaction. There have been many attempts to understand stock market behaviour and reduce the uncertainty in their predictions. We aim to analyse the influence and interaction amongst top performing companies' stock price and its social media exposure on Twitter. Specifically, we are going to present results on any causal relation between companies' data if it exists and discuss the relevance. To this end, several pre-processing transformations are applied on the data to suit the causal discovery algorithms. Moreover, we explore how time series data can be modelled either with time-delayed or instantaneous assumptions. Our study shows that linear correlation could be translated to conditional dependence relations and causal graphs. The code is available at `https://github.com/dssaenzml/ml703_timeSeriesCausalDiscovery`.

1 Introduction

Humans have a natural tendency to often think in terms of cause and effect i.e. if we understand why something happened, we change our behavior to improve future outcomes. Similarly, causality or causal inference enables machine learning algorithms to reason towards better predictions. Causal discovery is a subfield of causal inference which involves developing methods for learning cause and effect relationships from observational data. Knowledge of causal relations helps better understand a system and also provides critical insights about where a predictive model fails. On the other hand, time series data is a collection of observations that are generated through a period of time. It is created within various sectors such as financial firms, medical research and healthcare institutions, and many other organizations.

Stock price forecasting has been a popular problem related to time series. Most of the studies try to use any available information to predict future stock prices [**?**]. Dogan et al. [2020] attempts to find sentiment analysis on social media posts and then use this as a predictor for stock price in the future. However, no emphasis is usually made on investigating causes related to rise or fall of stock prices because of different interpretations on the causes. Understanding causal relationships can greatly reduce risk as well as increase returns in stock trading. Besides that, we are living in a technological era with heaps of instantaneous social media interaction. One natural question which arises is: can interactions on social networks help discover factors (if any) affecting stock performance for popular companies? Thus, our goal is to apply causal discovery on NASDAQ's stock prices performance for a few companies along with social media information from sources like Twitter.

1.1 Related work

There are several approaches to find the causal relation among variables. In the case of time series data, we have to be aware of the necessary conditions for it to be modelled as well as the assumptions.

In subsection 1.1.1, the previous works related to time series data are discussed. Causal discovery works are then summarized briefly in subsection 1.1.2.

1.1.1 Time series analysis

The analysis of time series information has been an ever growing subject for multiple fields. From finance to healthcare, time series analysis has shown several advancements in the reduction to uncertainty [Young and Shellswell, 1972, Dickey and Fuller, 1979, Zivot and Andrews, 1992]. In the case of finance, auto-regressive methods have had great acceptance because of its ability of using previous information for future predictions. These approaches rely heavily on stationary characteristics of the input data. Various transformations are performed to achieve stationary characteristics in the data. Decomposing the time series from its trend, seasonality, and cyclic behaviour usually helps to reduce the data into stationary [Dagum, 2013].

1.1.2 Causal discovery

Causal discovery involves the study of finding the possible cause and effect relationships within observed data. In recent years, causal discovery techniques have been applied to time series domains in order to discover factors responsible for a particular trend. Several causal discovery methods have been developed in general and also for time series in particular [Peters et al., 2012]. We describe some of the approaches which seem to be relevant to our problem statement. Granger causality is a very popular technique used for identifying causal relationships. It is based on the idea that if a factor Y 'Granger causes' a factor X, then the history of Y contains information that helps predict X above and beyond the information contained in past values of X alone. There are several methods based on Granger causality such as Vector Autoregressive models (VAR) Shojaie and Michailidis [2010] and Dynamic Time Warping (DTW) Amornbunchornvej et al. [2019].

Causal dependencies can also be found by testing conditional independence relations between variables and their pasts. Conditional independence based frameworks allow the causal graph identification under certain assumptions. Peter Clark (PC) [Spirtes and Glymour, 1991] developed a causal Markov condition and faithfulness based method for causal discovery. However, it is applied for the assumption that there is no confounder among variables. Fast causal inference (FCI) [Entner and Hoyer, 2010] is an approach that allows the possibility of taking latent confounding variables into account as opposed to methods based on Granger causality. Furthermore, there have been methods developed based on this idea such as PCMCI [Runge et al., 2019], PCMCI+ [Runge, 2020] which are variants of the popular PC algorithm adapted to time series data.

Besides Granger causality and conditional independence, structural equation models (SEM) have also been extensively used to perform causal discovery [Nogueira et al., 2021]. These kinds of models simultaneously account for measurement of error while examining the causal relationships. LiNGAM Shimizu et al. [2006] and TiMINo Peters et al. [2012] are some well known SEMs. Finally, there have been other methods based on score functions that are maximized such as in [Huang et al., 2018].

2 Dataset

The dataset that we have used in this study comes from Metin and Dogan [b,a] which was collected for social media sentiment analysis and stock performance for NASDAQ's top performing companies [Dogan et al., 2020]. The companies in this dataset are Apple (AAPL), Amazon (AMZN), Alphabet Inc Class C (GOOG), Alphabet Inc Class A (GOOGL), Microsoft Corporation (MSFT), and Tesla Inc (TSLA). Metin and Dogan [b] consists of daily number of posts,retweets of these posts, comments and number of likes for Twitter posts with the mentioned companies' hashtags on them. Due to the high correlation between the above mentioned features, we have selected the daily retweets number for each company on the tagged posts. Figure 3 shows that daily retweet numbers between companies are not highly correlated. On the other hand, Metin and Dogan [a] is composed of daily stock values, such as daily volume or high value, for the above companies where we are using close value only.

We have selected only business days of the NASDAQ stock market which excludes weekends as well as bank holidays. Moreover, we choose only the data that ranges from January 1st 2015 until December 31st 2019 such that there are 1304 data points for each feature. Since our dataset is time series, we decomposed them until they became stationary using the stationary tests which

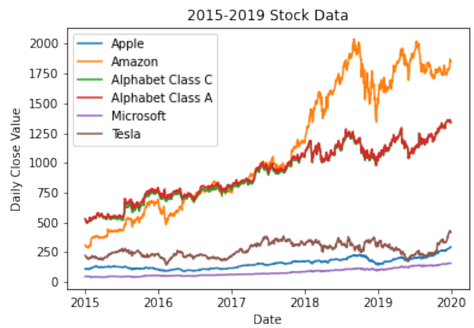

Figure 1: Daily close value price for the six top performing companies in NASDAQ from January 1st 2015 until December 31st 2019.

are explained in subsection 2.1. The different transformations done to the dataset are described in subsection 2.2.

2.1 Stationary test

Two widely used stationary tests are used to evaluate our time series data. First, we applied Augmented Dickey Fuller (ADF) Test as it helps to measure if the time series polynomial autoregressive lag has one root value equal to 1, or commonly known as unit root [Dickey and Fuller, 1979, Mushtaq, 2011]. If the time series can be explained with a unit root then its trend can be deviated at some point, meaning that the data distribution it explains could change at different periods of time. Therefore, if a time series does not have any root with unit value, it is considered as stationary by this test. However, this is a test with a low power for near-unit-root alternatives. So we decided to apply a second test to verify stationary of every time series.

Zivot-Andrews test [Zivot and Andrews, 1992] is another stationary test which also tries to reject the null hypothesis (a given time series is explained by a unit root in its autoregressive lag). Zivot-Andrews test is preferable for time series with structural break. Because of this reason, it was introduced in the case of oil price dramatic change, which could be favorable for the rapid volatile stock market data.

2.2 Data preprocessing

After having selected the daily retweets number and the daily close value difference for each of the six companies' stock markers, we applied various data transformations in order to find time series which pass the two mentioned stationary tests. For each of the companies' features, we applied logarithmic transformation, square root if non-positive values are present in the dataset. Then, to initial and newly-created features, we decomposed them by eliminating the trend on the time series. For the close value feature, we also applied one day difference to the initial values and the log-transformed values.

We performed tests on all resulting time series, initial and transformed ones, with the ADF and Zivoc-Andrews test such that we filtered all those time series features that pass these two stationary tests. As it is shown in Figure 2 and 3, we selected the daily close value difference for all companies, and the log and square transformed daily retweets number for five of the six companies. For the special case of Tesla-related daily retweets number, we used the detrended log transformed value since this was the only time series for this company that passed both stationary tests.

On one hand, in Figure 2 we can observe the linear correlation coefficient value among the six companies lagged daily close value. Alphabet Inc Class C (GOOG) and Alphabet Inc Class A (GOOGL) are the most correlated time series for stock close values as expected, i.e. those two NASDAQ's markers come from the same holding company Alphabet Inc. The second highest linear correlation value is between Amazon and Microsoft Corporation which is corroborated by Figure 4 **(A)** scatter plot. On the other hand, we can see from the low correlation coefficient and Figure 4 **(C)**, that the daily retweets numbers are not very much linearly related to each other. Cross-relation

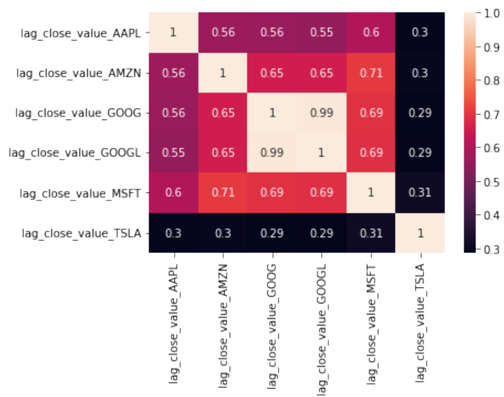

Figure 2: Correlation matrix for company related stock return data.

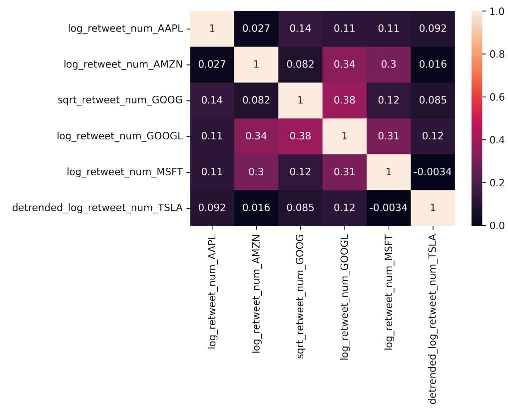

Figure 3: Correlation matrix for company related retweet data.

(A) **(B)** **(C)**

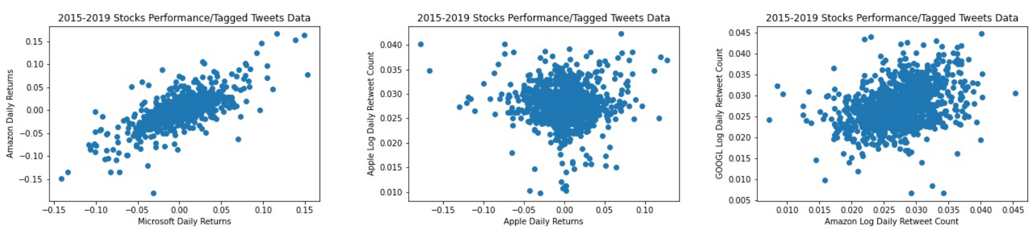

Figure 4: **(A)** Scatter plot for AMZN and MSFT daily close value difference, i.e. $x_t - x_{t-1}$. **(B)** Scatter plot for AAPL daily close value difference vs AAPL log retweet number. **(C)** Scatter plot for AMZN log retweet number vs GOOGL log retweet number.

between close value and retweets number data for all companies is very non-linear as seen in the linear correlation coefficient plot in Figure S1 and the scatter plot for Apple data in Figure 4 **(B)**.

3 Methods

We use Python-language-based libraries on multivariate causal discovery. The two packages that we have studied are Tigramite Runge, and PyTrad [Zhang et al., 2021]. To test data's conditional independence, we investigated with different statistical significance values for conditional independence tests. In addition, after having preprocessed our variables, we understood the linear and non-linear relation among them. This influences our choice of conditional independence test and their respective results.

Using Tigramite, we applied their implementation of Peter Clark causal discovery with Momentary Conditional Independence tests (PCMCI) [Spirtes and Glymour, 1991, Runge et al., 2019, Runge, 2020] with two different conditional test: partial correlation test [Lawrance, 1976], and Gaussian Process regression with Distance Correlation test on the residuals (GPDC) [Székely et al., 2007]. Tigramite and its implementations were built under the assumption of time-delayed causal relations between variables.

The second package is Causal-Learn, formerly known as Pytrad, [Zhang et al., 2021], which works with instantaneous causal relations between variables. We applied three causal discovery algorithms from this library: Peter Clark (PC) algorithm [Spirtes and Glymour, 1991], Fast Causal Inference (FCI) [Spirtes et al., 1995], and Greedy Equivalence Search (GES) [Chickering, 2002a, Huang et al., 2018]. For the former two algorithms we used two conditional independence tests: Fisher-Z's test [Fisher, 1915], and Kernel-based Conditional Independence (KCI) test [Zhang et al., 2012]. For the latter one, we used generalized score [Huang et al., 2018]. We discuss these algorithms and tests in more detail in the following subsections.

3.1 PCMCI Algorithm

PCMCI is a causal discovery method specific to time series data. It consists of two steps - (1) PC algorithm that removes irrelevant conditions for all variables by iterative independence testing and (2) momentary conditional independence (MCI) test which reduces false positives for the highly-interdependent time series.

Partial Correlation test:

Partial correlation quantifies the correlation between two variables when conditioned on one or several other variables. Hence, when there is a correlation between two variables, it might be partially explained by a confounder.

Gaussian Process regression with Distance Correlation test:

Gaussian process regression is a non-parametric method which calculates the posterior using the training data and computes the predictive posterior distribution on test data. In addition to this, distance correlation measures both linear and nonlinear association between two random variables.

3.2 PC Algorithm

The PC algorithm is a constraint based method for causal discovery. It is based on causal Markov condition and faithfulness assumption, when there is no confounder (unobserved direct common cause of two measured variables) [Spirtes and Glymour, 1991]. The algorithm starts with a fully connected graph and tries to prune the graph based on unconditional and conditional independence relations.

3.3 FCI Algorithm

Fast causal inference (FCI) algorithm is based on PC algorithm but is capable of discovering unknown confounding variables. As PC, FCI also uses statistical independence relations to prune an undirected graph. So it builds correct graphs if the causal Markov assumption, the faithfulness assumption, the assumptions of no selection bias, and that independence relations can be reliably tested [Spirtes et al., 1995].

3.4 GES Algorithm

As the name suggests, Greedy Equivalence Search (GES) is a greedy algorithm that searches over equivalence classes of Directed Acyclic Graphs (DAGs). Greedy search proceeds at each step by evaluating each neighbor of the current state, and moving to the one with the highest score if doing so improves the score. The set of neighbors of each state in the search defines the connectivity of the search space Chickering [2002b]. GES starts with an empty graph and since its greedy, temporarily adds all required edges and tries to remove unnecessary edges as it progresses. It decides to add a directed edge in an attempt to increase fit measured by some score such as BIC, marginal likelihood etc. Thus, the edge that most improves fit is added. Once the score no longer improves, the GES starts removing edges to improve the score. We have used a generalized score with cross validation for data with single-dimensional variates.

4 Results

We present some results on the methods mentioned above with some specific characteristics per algorithm.

4.1 PCMCI

For our experiments on time-delayed interaction assumption, we set the significant statistical threshold value very conservative given the multiple causal relations that may appear with previous time-steps

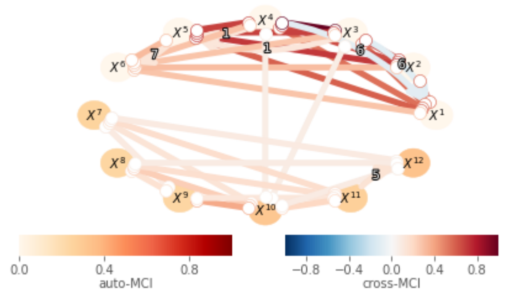

Figure 5: PCMCI algorithm with Partial Correlation conditional independence test [Runge et al., 2019, Runge, 2020]. (Refer to Table S1 for full variable names)

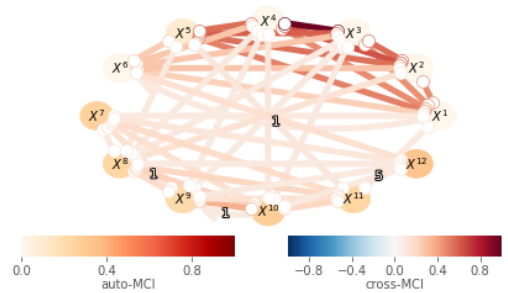

Figure 6: PCMCI algorithm with GPDC conditional independence test [Székely et al., 2007]. (Refer to Table S1 for full variable names)

for a variable in itself and in relation to other variables and previous time-steps. So we chose initially an alpha value of 0.01 for rejecting conditional dependence between data features. We may see the result of using PCMCI with our 12 features for the six companies using the two previously described conditional independence test in Figure 5 and 6. Given the large amount of features to analyse and regression relations up to 7 days for all of them, both graphs show many causal links. The most interesting results lie on the linear separation present in the partial correlation conditional independence test. Though, it can be seen that the MCI, as well as self-MCI, strength between most variables is close to 0. Finally, Figure 5 shows a causal link up to the seventh day between companies' return values.

4.2 PC

The PC algorithm performance on our preprocessed dataset presented great causal graphs. We check linear and non-linear methods for the conditional independence test which produce similar results as can be seen in Figure 7 and 8. As mentioned previously, the non-linear method for conditional independence, KCI, was used by applying a Gaussian kernel for data transformation. This experiment showed similar results to the linear case, although it produced interesting interactions with the social interaction for some of the companies' return value variables. An interesting difference between the two methods are the direction of the causal relations between stock returns, e.g. Apple's return seems to be the end-point for KCI's results, though for Fisher-Z's framework it was the root. In both graphs, green links are the representation for undirected relations between two variables and it is present for the same two variables GOOG and GOOGL return values. All conditional independence tests performed using the PC algorithm were done with a 0.05 threshold for rejecting conditional independence between variables.

4.3 FCI

Similarly, we used the FCI algorithm in order to find causal links. Likewise, we used two conditional independence tests to observe linear and non-linear conditions applied to it - Fisher-Z's and Kernel based Conditional Independence test (KCI) respectively. For this algorithm that assumes instantaneous interaction between variables, we set the alpha value for conditional independence test to 0.05 to allow slightly more causal relations to be found. From Figure 9, it can be clearly observed that there is no causal relation between company close values and retweets. We also notice that there are causal relations among different company close values. Similarly, there are causal relations among company retweets as well. Edges in the graph have several representations: one with blank circles at either ends meaning that there is no clear definition of the direction towards a node, arrows at the ends of edges representing a found causal relation, and a simple line at the extreme meaning that there was no direction found. From Figure 10 for FCI with KCI based test, we obtain several disjoint graphs. However, there are few causal relations that have been found between Amazon, Alphabet A and Microsoft close values.

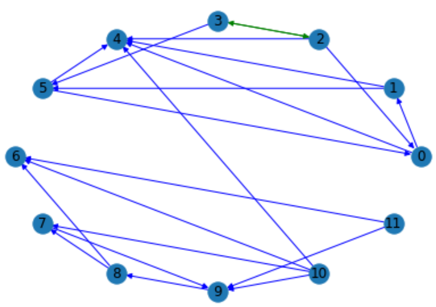

Figure 7: PC algorithm with Fisher-Z's conditional independence test [Fisher, 1915]. (Refer to Table S1 for full variable names)

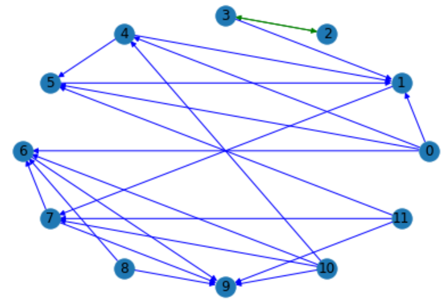

Figure 8: PC algorithm with KCI's conditional independence test with Gaussian kernel [Zhang et al., 2012]. (Refer to Table S1 for full variable names)

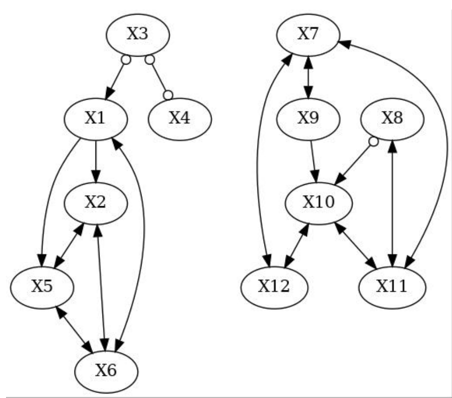

Figure 9: FCI algorithm with Fisher-Z's conditional independence test [Fisher, 1915]. (Refer to Table S1 for full variable names)

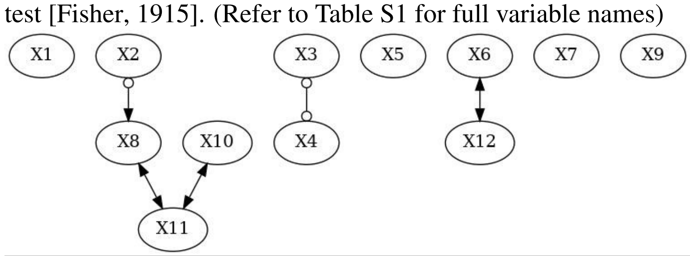

Figure 10: FCI algorithm with KCI's conditional independence test with Gaussian kernel [Zhang et al., 2012]. (Refer to Table S1 for full variable names)

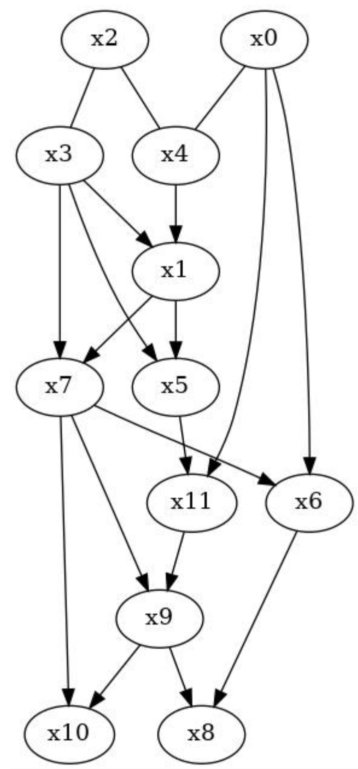

Figure 11: GES algorithm with general score function [Chickering, 2002a, Huang et al., 2018]. (Refer to Table S1 for full variable names)

KCI tests found some interesting relations in the case of FCI algorithm. For example, it found that Amazon daily stock returns are conditionally independent to Apple-related daily retweet counts given Alphabet Class A (GOOGL) stock returns with 0.22 for a the mentioned threshold of $p = 0.05$. This translated into disconnected relations among the conditionally independent variables. Beyond the causal graph, this result from the KCI test does not seem to be an immediate relation because of the complex information exchange amongst companies' variables. On the other hand, a more intuitive result for KCI test was that of Apple and Alphabet class A stock returns conditional independence given Alphabet class C stock return (0.79 significant value).

4.4 Greedy Equivalence Search (GES)

The graph obtained from the GES algorithm shows two kinds of edges i.e., directed and undirected edges. The undirected edges represent the existence of some relation (correlation) between the connected nodes whereas the directed edges represent the causal relation between the connected nodes.

Analysing stock price relations: We can interpret from the Figure 11 that there is a relationship between all the companies stock prices. Alphabet Class C stock price has correlation with Microsoft and Alphabet Class A stock price. Similarly, Apple stock price shows correlation with Microsoft stock price. There is a causal chain between Alphabet Class A, Amazon and Tesla stock prices.Similarly, there is a causal chain between Microsoft, Amazon and Tesla stock prices. Alphabet Class A and Microsoft stock prices appear to be common causes affecting Amazon and Tesla stock prices.

Analysing company tweets relations: From the Figure 11, we observe that there are only causal relationships between companies' retweets. Amazon retweets frequency affects Microsoft, Alphabet Class A and Apple retweets. However, there are no company retweets influencing Amazon.Alphabet Class A acts a common cause for Alphabet Class C and Microsoft retweets. Apple retweets appear to affect only Alphabet Class C. Amazon and Tesla retweets are affecting Alphabet Class A retweets.

Analysing relations between company stock prices and retweets frequencies: It can be observed from the Figure 11 that there exist causal relations between company stock price and retweets. Alphabet Class A stock price appears to affect all companies' retweets directly or indirectly. Interestingly, Microsoft stock price does not have a direct causal effect on its retweets. Rather, a long causal chain is observed through different companies such as Amazon, Tesla, Alphabet Class A. Apple stock price has influence on all company retweets except Amazon. We observe that there are plenty of causal relations from stock price to retweets frequency. *However, there are no causal relations from retweets to stock prices which seems to be reasonable.*

5 Discussion

The assumption of time-delayed influence on daily stock pricing did not hold very well given the weak causal relations found with the PCMCI algorithm, see Figure 5 and 6. Applying linear and nonlinear conditional independence tests for PCMCI algorithm did not produce stronger links among the companies' variables, or its seven respective regressive counter-parts. Moreover, there are no clear directions for any of the links. Thus, given this result we decided to look into instantaneous methods for causal discovery for our dataset.

The resulting causal graphs for PC and FCI algorithms with Fisher-Z's conditional independence test reflect the linear correlations found in section 2.2 and depicted in Figure S1, as seen in Figure 7 and 9. Most of the relations found by these two methods are the same with the exception of a few like the connection between Alphabet Class A class value return and that of Tesla's. One clear causal relation discrepancy between these two is the link between Microsoft retweet number and its close value return. This relation was only found by the PC algorithm and it is the only link between the two types of company data. Conversely, the causal graphs found by PC and FCI algorithms with the non-linear conditional independence test, KCI, resulted in two completely different graphs. Figure 8 shows a similar graph to that of the PC Fisher-Z's conditional test with extra found links between daily retweet number and close value return from all six company stock markers. There are causal connections from Tesla's daily retweet number to close value return. Microsoft's variables are also related similarly in Figure 8, showing some intuitive relations from instantaneous social media interaction and end of day stock price value. On the other hand, FCI algorithm with KCI test performs poorly on our dataset as it can be seen in Figure 10. Four variables are completely disconnected to any other input feature. The only found relations are between the same companies' variables for Amazon, Alphabet's two markers, and Tesla. GES algorithm identifies relations among different entities as shown in Figure 11. Similar to PC and FCI algorithms, GES also shows relations among all company stock prices. Likewise, there are also causal relations found among the company retweets as well. There are also several causal relations observed from stock prices to retweets which is sensible. Furthermore, GES shows no causal relations from retweets to stock prices which seems to be logical and is not observed with the other algorithms like PC and FCI.

6 Conclusion

In our work we observed the application of causal discovery on time series information. Time series have to be transformed into readily stationary data in order to reduce the influence of random noise for further analysis. We achieve this with several mathematical transformations, stationary tests, and linear correlation analysis for all observed data. Given all described pre-processing, we analysed how different assumptions on the data demonstrated different behaviour on causal graphs. Initially we assumed there could be lagged-relations between the different variables. This resulted in very complicated graphs with low causal strength in most of the relations. Our following assumption on instantaneous information relation provided clearer and more reasonable causal links for the observed information. Linear correlation also has shown influence in the resulting graphs that applied linear-based conditional independence test. We believe that causal discovery analysis such as in this study could improve stock market understanding and predictability.

Acknowledgement
We are thankful to our professor Dr.Kun Zhang for his valuable feedback.

References

C. Amornbunchornvej, E. Zheleva, and T. Y. Berger-Wolf. Variable-lag granger causality for time series analysis. *2019 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, Oct 2019. doi: 10.1109/dsaa.2019.00016. URL `http://dx.doi.org/10.1109/DSAA.2019.00016`.

D. M. Chickering. Optimal structure identification with greedy search. *J. Mach. Learn. Res.*, 3: 507–554, 2002a.

D. M. Chickering. Optimal structure identification with greedy search. *Journal of machine learning research*, 3(Nov):507–554, 2002b.

E. Dagum. Time series modelling and decomposition. *Statistica*, 70, 05 2013. doi: 10.6092/issn.1973-2201/3597.

D. Dickey and W. Fuller. Distribution of the estimators for autoregressive time series with a unit root. *JASA. Journal of the American Statistical Association*, 74, 06 1979. doi: 10.2307/2286348.

M. Dogan, O. Metin, E. Tek, S. Yumuşak, and K. Öztoprak. Speculator and influencer evaluation in stock market by using social media. In *2020 IEEE International Conference on Big Data (Big Data)*, pages 4559–4566, 2020. doi: 10.1109/BigData50022.2020.9378170.

D. Entner and P. Hoyer. On causal discovery from time series data using fci. 2010.

R. A. Fisher. Frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population. *Biometrika*, 10(4):507–521, 1915. ISSN 00063444. URL `http://www.jstor.org/stable/2331838`.

B. Huang, K. Zhang, Y. Lin, B. Schölkopf, and C. Glymour. Generalized score functions for causal discovery. *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2018.

A. J. Lawrance. On conditional and partial correlation. *The American Statistician*, 30(3):146–149, 1976. ISSN 00031305. URL `http://www.jstor.org/stable/2683864`.

O. Metin and M. Dogan. Values of top nasdaq companies from 2010 to 2020. Available at `https://www.kaggle.com/omermetinn/values-of-top-nasdaq-copanies-from-2010-to-2020` (2010-2020), a.

O. Metin and M. Dogan. Tweets about the top companies from 2015 to 2020. Available at `https://www.kaggle.com/omermetinn/tweets-about-the-top-companies-from-2015-to-2020` (2015-2020), b.

R. Mushtaq. Augmented dickey fuller test. *Econometrics: Mathematical Methods & Programming eJournal*, 2011.

A. R. Nogueira, J. Gama, and C. A. Ferreira. Causal discovery in machine learning: Theories and applications. *Journal of Dynamics and Games*, 8(3):203–231, 2021.

J. Peters, D. Janzing, and B. Schölkopf. Causal inference on time series using structural equation models, 2012.

J. Runge. Tigramite – causal discovery for time series datasets. Available at `https://github.com/jakobrunge/tigramite` (2014-2020).

J. Runge. Discovering contemporaneous and lagged causal relations in autocorrelated nonlinear time series datasets, 2020.

J. Runge, P. Nowack, M. Kretschmer, S. Flaxman, and D. Sejdinovic. Detecting and quantifying causal associations in large nonlinear time series datasets. *Science Advances*, 5(11):eaau4996, Nov 2019. ISSN 2375-2548. doi: 10.1126/sciadv.aau4996. URL `http://dx.doi.org/10.1126/sciadv.aau4996`.

S. Shimizu, P. O. Hoyer, A. Hyvärinen, and A. Kerminen. A linear non-gaussian acyclic model for causal discovery. *J. Mach. Learn. Res.*, 7:2003–2030, Dec. 2006. ISSN 1532-4435.

A. Shojaie and G. Michailidis. Discovering graphical granger causality using the truncating lasso penalty, 2010.

P. Spirtes and C. Glymour. An algorithm for fast recovery of sparse causal graphs. *Social science computer review*, 9(1):62–72, 1991.

P. L. Spirtes, C. Meek, and T. S. Richardson. Causal inference in the presence of latent variables and selection bias. In *UAI*, 1995.

G. J. Székely, M. L. Rizzo, and N. K. Bakirov. Measuring and testing dependence by correlation of distances. *The Annals of Statistics*, 35(6), Dec 2007. ISSN 0090-5364. doi: 10.1214/009053607000000505. URL `http://dx.doi.org/10.1214/009053607000000505`.

P. Young and S. Shellswell. Time series analysis, forecasting and control. *IEEE Transactions on Automatic Control*, 17(2):281–283, 1972. doi: 10.1109/TAC.1972.1099963.

K. Zhang, J. Peters, D. Janzing, and B. Schoelkopf. Kernel-based conditional independence test and application in causal discovery, 2012.

K. Zhang, J. Ramsey, S. Shimizu, P. Spirtes, and C. Glymour. Pytrad, 2021. URL `https://github.com/cmu-phil/pytrad`.

E. Zivot and D. W. K. Andrews. Further evidence on the great crash, the oil-price shock, and the unit-root hypothesis. *Journal of Business & Economic Statistics*, 10(3):251–270, 1992. ISSN 07350015. URL `http://www.jstor.org/stable/1391541`.

Supplementary Material

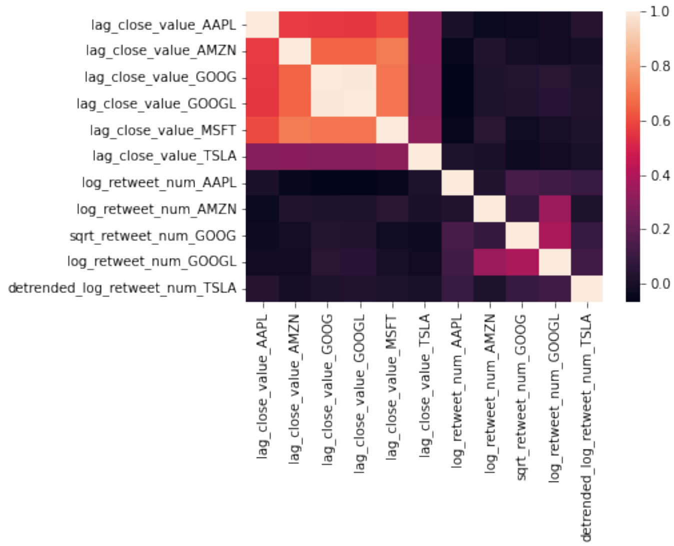

Figure S1: Cross-data linear correlation coefficients matrix for all six companies NASDAQ's markers.

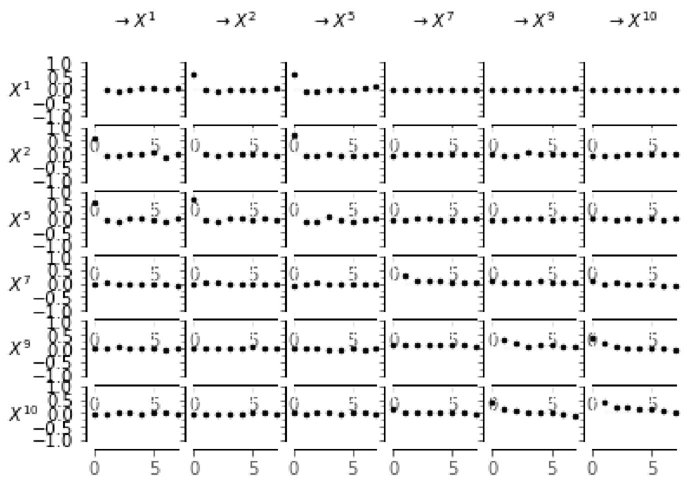

Figure S2: Auto-regressive correlation plot for several variables as labelled by Table S1. As it can be seen, there is not any apparent auto-correlation to previous time-steps for most variables beyond $t - 3$.

Table S1: Legend for variables in Figures.

Variable	**Labels for Causal Discovery (PC / FCI / GES)**
1-day-difference Close value of Apple (AAPL)	0 / X1 / X0
1-day-difference Close value of Amazon (AMZN)	1 / X2 / X1
1-day-difference Close value of Alphabet Class C (GOOG)	2 / X3 / X2
1-day-difference Close value of Alphabet Class A (GOOGL)	3 / X4 / X3
1-day-difference Close value of Microsoft (MSFT)	4 / X5 / X4
1-day-difference Close value of Tesla (TSLA)	5 / X6 / X5
Logged Retweet number of Apple (AAPL)	6 / X7 / X6
Logged Retweet number of Amazon (AMZN)	7 / X8 / X7
Squared Retweet number of Alphabet Class C (GOOG)	8 / X9 / X8
Logged Retweet number of Alphabet Class A (GOOGL)	9 / X10 / X9
Logged Retweet number of Microsoft (MSFT)	10 / X11 / X10
Detrended logged Retweet number of Tesla (TSLA)	11 / X12 / X11