Non-parametric Causality Detection: An Application to Social Media and Financial Data

Fani Tsapeli
School of Computer Science
University of Birmingham
Edgbaston B15 2TT
Birmingham, UK
t.tsapeli@cs.bham.ac.uk

Mirco Musolesi
Department of Geography
University College London
Gower Street WC1E 6BT
London, UK
m.musolesi@ucl.ac.uk

Peter Tino
School of Computer Science
University of Birmingham
Edgbaston B15 2TT
Birmingham, UK
p.tino@cs.bham.ac.uk

ABSTRACT
According to behavioral finance, stock market returns are influenced by emotional, social and psychological factors. Several recent works support this theory by providing evidence of correlation between stock market prices and collective sentiment indexes measured using social media data. However, a pure correlation analysis is not sufficient to prove that stock market returns are influenced by such emotional factors since both stock market prices and collective sentiment may be driven by a third unmeasured factor. Controlling for factors that could influence the study by applying multivariate regression models is challenging given the complexity of stock market data. False assumptions about the linearity or non-linearity of the model and inaccuracies on model specification may result in misleading conclusions.

In this work, we propose a novel framework for causal inference that does not require any assumption about a particular parametric form of the model expressing statistical relationships among the variables of the study and can effectively control a large number of observed factors. We apply our method in order to estimate the causal impact that information posted in social media may have on stock market returns of four big companies. Our results indicate that social media data not only correlate with stock market returns but also influence them.

Categories and Subject Descriptors
G.3 [Probability and Statistics]: Time-series Analysis

General Terms
Causal Inference, Time-series analysis

Keywords
Causality, social media, stock market, sentiment tracking, time-series

1. INTRODUCTION
We are living in the era of social media, using tools such as Facebook, Twitter and blogs to communicate with our friends, to share our experiences and to express our opinion and emotions. Recently, mining and analyzing this kind of data has emerged as an area of great interest for both the industrial and academic communities. Several studies have examined the ability of social media to serve as crowdsensing platforms. For example, authors in [1] demonstrate that social media can monitor the popularity of products or services and predict their future revenues. Evidence has been found that social media can be used to predict election results [2] or even stock market prices [3].

Most studies so far have focused on using social media data as early indicators of real-world events. But to what extent do opinions expressed through social media actually have a causal influence on the examined events? For example, are stock market prices influenced by the opinions and sentiments that are reported in social media, or is it the case that stock market prices and sentiments are driven only by other (e.g. financial) factors? Would the results have been different if we could manipulate social media data? In order to answer such questions a causality study is required.

Some recent studies have examined the ability of social media to influence real-world events by applying randomized control trials. For example, authors in [4] examine the effect of political mobilization messages by using Facebook to deliver such messages to a randomly selected population; the effect of the messages is measured by comparing the real-world voting activity of this group with the voting activity of a control group. Similarly, in [5] authors use randomized trials in order to examine the social influence of aggregated opinions posted in a social news website. Randomized control trials are a reliable technique for conducting causal inference studies [6]. However, their applicability is limited since they require scientists to gather data using experimental procedures and do not allow the exploitation of the large amount of observational data. In many cases, it is not feasible to apply experimental designs or it is considered unethical.

In this work, we study the causal impact of social, psychological and emotional factors on stock market prices of big companies using observational data collected through Twitter. Twitter enables us to capture people sentiments and opinions about traded assets and their reactions on related news and events. Previous works have demonstrated that social media data correlate with stock market prices [7-9].
These studies were predominantly based on correlation or Granger causality analysis. Granger causality tests the ability of a time-series to predict values of another one \[10\]. However, it cannot be used to discover real causality. A positive result on a Granger causality test does not necessarily imply that there is a causal link between the examined time-series since both the examined time-series may be influenced by a third variable (confounding bias). Multivariate regression techniques can be applied in order to control for confounding bias. Some studies attempt to improve the accuracy of stock market prediction models by applying multivariate regression \[11\]. However, the focus of these works is on prediction rather than on causal inference. Applying regression models for causal inference suffers from two main limitations. First, stock market prices can be influenced by a large number of factors such as stock market prices of other companies \[12,13\], foreign currency exchange rates and commodity prices. Such factors may also influence people sentiments. Consequently, to eliminate any confounding bias one is required to include a large number of predictors in the regression model. Estimation of regression coefficients in a model with a large number of predictors can be challenging. When data dimensionality is comparable to the sample size, noise may dominate the ‘true’ signal, rendering the study infeasible \[14\]. Second, inaccuracies in model specification, estimation or selection may result in invalid causal conclusions.

Given the limitations of parametric methods, we propose a novel framework for causal inference in time-series that is based on matching design \[15,16\]. This technique attempts to eliminate confounding bias by creating pairs of similar treated and untreated objects, i.e. objects with similar values on baseline characteristics that could influence the causality study. Thus, the effect of an event is estimated by comparing each object exposed to an event with a similar object that has not been exposed. Matching design bypasses the limitations of regression-based methods since it does not require specification of a model class. However, it cannot be applied in time-series since it assumes that the objects of the study are realizations of i.i.d variables. We reformulate the concept of matching design to make it suitable for causal inference on time-series data. In our case the time-series collection includes treatment time-series \(X\), response time-series \(Y\) and a set of time-series \(Z\) which contain characteristics relevant to the study. The units of our study correspond to time-samples; the \(t^{th}\) unit is characterized by a treatment value \(X(t)\), a response value \(Y(t)\) and a set of values representing baseline characteristics \(Z(t)\). We assess the causal impact of a time-series \(X\) on \(Y\) by comparing different units (i.e. time-samples) on \(Y\) after controlling for characteristics captured in \(Z\). As explained in Section 2 we methodology assures that the objects are uncorrelated, which is a weaker version of the independence assumption requirement of the matching design. We apply our framework in order to estimate the causal impact that the sentiment of information posted in social media may have on traded assets. In detail, we estimate a daily sentiment index (treatment time-series) based on information posted in Twitter and we assess its impact on daily stock market closing prices (response time-series) of four big technological companies after controlling for other factors (set of time-series \(Z\)) that may influence the study, such as the performance of other big companies.

In summary, the contribution of this work is twofold:

1. We propose a causal inference framework for time-series that can be applied to high-dimensional data without imposing any restriction on the model class describing the associations among the data. We demonstrate, using synthetic data, that our methodology is more effective on detecting true causality compared to other methods that have been applied so far, for causal inference in time-series.

2. We apply our method in order to quantify the causal impact of emotional and psychological factors, captured by social media, on stock market prices of four technological companies. To the best of our knowledge, this is the first study that measures the causal influence of such factors on finance. It should be noted that, since all the examined companies belong to the technological sector, our findings cannot be directly generalized for any company.

The rest of this paper is organized as follows. In Section 2 we discuss the main methodologies that are used for causal inference. In Section 3 we present the proposed framework. In Section 4 we evaluate our approach on synthetic data, in conjunction with other methods that have been previously applied for causal inference in time-series. Moreover, we apply our method in order to assess the causal impact of information posted in Twitter on stock market prices of specific companies. In Section 5 we discuss some relevant works which attempt to uncover the relationship between social media data and stock market movement. In Section 6 we summarize the advantages and limitations of the proposed approach and we discuss its computational cost. Finally, Section 7 concludes the paper by summarizing our contributions.

2. BACKGROUND ON CAUSAL INFERENCE

Causal analysis attempts to understand whether differences on a specific characteristic \(Y\) within a population of units are caused by a factor \(X\). \(Y\) is called response, effect or outcome variable and \(X\) treatment variable or cause. Units are the basic objects of the study and they may correspond to humans, animals or any kind of experimental objects.

2.1 Potential Outcome Framework

The key concept on causation theory is that given a unit \(u\), the value of the corresponding response variable \(Y(u)\) can be manipulated by changing the value of the treatment variable \(X(u)\) \[17,18\]. In this paper, we will consider \(X(u)\) as a binary treatment variable. Hence, there will be two treatments: \(x_1\) for treated units and \(x_0\) for untreated. We denote by \(Y_1(u)\) the value of \(Y\) when \(X(u) = x_1\) and \(Y_0(u)\) when \(X(u) = x_0\). In order to test the effect of \(x_1\) on unit \(u\), we need to estimate the quantity \(Y_1(u) - Y_0(u)\). The fundamental problem of causal inference is that it is impossible to observe both \(Y_1(u)\) and \(Y_0(u)\) on the same unit \(u\) and, therefore, it is impossible to measure the real causal effect of \(x_1\) on the unit. \[17,18\]. Thus, the average effect of a treatment \(X\) is estimated by comparing a population of objects that received the treatment \(x_1\) with a population that received the treatment \(x_0\) and evaluating the corresponding average values of the effect variable \(Y\). We denote by \(Y_1\) and \(Y_0\) random variables representing the outcome variable when the treatments \(x_1\) and \(x_0\) are applied, respectively. We also define a (random) variable \(DY = Y_1 - Y_0\). Then, the average
treatment effect (ATE) of $x_1$ is estimated as the expected value $E(DY)$.

The average treatment effect can be estimated only if the following three assumptions are satisfied:

1. the effect differences are i.i.d. realizations of $DY$.

2. the observed outcome in one unit is independent from the treatment received by any other unit (Stable Unit Treatment Value Assumption - SUTVA).

3. the assignment of units to treatments is independent from the outcome (ignorability). Ignorability can be formally expressed as $Y_1 \perp X, Y_0 \perp X$. The assumption of ignorability requires that all the units have equal probability to be assigned to a treatment. If this assumption does not hold, the units that received a treatment $x_1$ may systematically differ from units that did not receive such a treatment. In such a case the average value of the outcome variable of the treated units could be different from that of other units, even if the treatment had not been received at all.

In experimental studies, ignorability can be achieved by randomly assigning units to treatments. However, in observational studies this is not feasible. Instead, the average treatment effect can be estimated by relaxing ignorability to conditional ignorability. According to conditional ignorability assumption, the treatment assignment is independent from the outcome, conditional on a set of confounding variables $H$. Variables $H$ represent baseline characteristics of the units that are considered relevant for the study (e.g. in a medical study that examines the impact of a drug, baseline characteristics could be the previous health condition of the units (in this case patients), their age etc.). Thus, conditional ignorability is expressed as $Y_1 \perp X|H, Y_0 \perp X|H$. The variables that must be included in the set $H$ in order to achieve conditional ignorability are also called confounding variables. Conditional ignorability cannot be achieved when one or more confounding variables are unobserved. The main limitation of all non-experimental causality studies is that the possibility that important confounding variables are missing cannot be eliminated. Latent variable models [19] and analysis on the sensitivity of the conclusions on missing confounding variables [20] have been proposed to handle this issue.

There are two main methodologies that are applied in order to achieve conditional ignorability: regression and matching [15]. Regression expresses the outcome variable $Y$ as a function of the treatment variable $X$ and the set of variables $H$ [21,22]. Linear models are usually applied. Methods based on regression require scientists to correctly specify a regression model. These methods are affected by the typical problems of parametric approaches to causality detection (i.e. model misspecification, overfitting and poor performance on high-dimensional datasets).

Matching comprises a more flexible methodology for causal inference in observational data since it does not require the specification of a model class [16]. Conditional ignorability is achieved by creating sub-population within which the values of the confounding variables $H$ are the same or similar. Thus, considering a set $G$ of pairs of treated and untreated (control) units $(u,v)$ such that $H(u) \approx H(v)$, we can estimate the average treatment effect (ATE) as

$$
\hat{E}(DY) = \frac{\sum_{(u,v) \in G} Y(u) - Y(v)}{|G|},
$$

where $|G|$ denotes the size of $G$. Scientists need to assess the degree of similarity between the matched treated and control units. Similarity relation ($\approx$) can be assessed by estimating the standardized mean difference for each confounding variable between matched treated and control units, or by applying graphical methods such as quantile-quantile plots, cumulative distribution functions plots, etc. [23,26]. If sufficient balance has not been achieved, the applied matching method needs to be revised.

### 2.2 Directed Acyclic Graphs

Pearl [27,29] proposed the use of directed acyclic graphs (DAGs) for representing causal relationships. In causal graphs, nodes represent the variables of the experiment. If $P$ a set of the direct predecessors (parents) of a node $Y$, a direct arrow from a node $Q$ (can represent $X$ or one of the variables of set $Z$) to $Y$ will exist only if $Y \not\perp Q|P \setminus \{Q\}$. A direct arrow from a node $Q$ to a node $Y$ represents a causal relationship between the two variables (i.e. $Q$ causes $Y$).

Pearl also introduces a graphical criterion for defining a sufficient set of variables that need to be controlled in order to achieve conditional ignorability when testing the causal impact of a variable $X$ on a variable $Y$ (back-door criterion). According to this rule, a subset $H$ of variables is sufficient if no element of $H$ is a descendant of $X$ and the elements of $H$ block all paths from $X$ to $Y$ that end with an arrow to $X$ (back-door paths). The intuition behind this criterion is that back-door paths from $X$ to $Y$ represent spurious associations and therefore need to be excluded in order to obtain unbiased estimation of the causal effect of $X$ on $Y$.

Causal graphs differ from Bayesian graphs since, in a causal graph, an arrow from $Q$ to $Y$ denotes that $Y$ values would change if we could manipulate the values of $Q$. However, this hypothesis cannot be tested in observational studies. Instead, a graph is created either by utilizing prior knowledge about the structure of the model or by conducting conditional independence tests on the observational data [30–33]. The correctness of the graph can be assessed by fitting the observational data to a system of structural equations derived from the graph (i.e. each variable is regressed on all its direct predecessors) [29].

### 2.3 Causality on Time-series

Causality studies on time-series have been largely based on Granger causality [34]. The Granger causality test examines if past values of one variable are useful in prediction of future values of another variable. In detail, a time-series $X$ Granger causes a time-series $Y$ if modeling $Y$ by regressing it on past values of both $Y$ and $X$ results in reduced residual noise compared to a simple autoregressive model. Granger causality does not test real causality since the conditional ignorability assumption is not satisfied, i.e., the values of both treatment variable $X$ and control variable $Y$ may be driven by a third variable. Moreover, it considers only linear relationships. Granger causality has been extended to handle multivariate cases [35] as well as non-linear cases [36,37].

However, as it was mentioned also at the introduction, inaccuracies on model specification may result in misleading conclusions. In [38] the authors propose an additional model
3. PROPOSED MECHANISM

Given the previously discussed limitations on existing methodologies for causality discovery in time-series, we propose a novel framework that enables causal inference in time-series data that is based on matching design and therefore does not require specification of a model class. The proposed framework also requires only few conditional independence tests, thus it can handle more effectively high-dimensional data. Denote by \( Y \) the effect and by \( X \) the cause, respectively (i.e., if \( X(t^e) \) the \( i \)-th sample of \( X \), \( X(t^e) = X(t^e_{i-1}) \)). At Figure 1 we provide a graphical representation of time-series \( X \) along with its first \( L \) lagged versions.

For example, consider a study that examines the effects of an industry on the pollution level in a region based on weekly measurements. In this case, a unit of the study corresponds to one week and the variables of the study to weekly pollution measurements, industrial waste and other relevant characteristics. By Figure 2 we graphically depict the notion of a unit in our time-series matching design framework in comparison with the traditional notion of unit on causality studies. In the rest of this paper, the terms 'unit' and 'time-sample' will be used interchangeably.

In order to build a graph, we examine the dependencies between the variables \( X, Y, Z \) and all the other variables of the set \( S \). In order to examine if two time-series \( X \) and \( Y \) are independent (assuming that the time-series are stationary in the first two moments) we can estimate the Pearson correlation coefficient as follows:
\[ r_{xy} = \frac{\sum_{u=0}^{N}(X(t_u^x) - \bar{X})(Y(t_u^y) - \bar{Y})}{\sqrt{\sum_{u=0}^{N}(X(t_u^x) - \bar{X})^2 \sum_{u=0}^{N}(Y(t_u^y) - \bar{Y})^2}} \]

with \( \bar{X}, \bar{Y} \) the sample means of \( X, Y \) respectively. Vanishing correlation could be considered as indication of independence between the examined time-series. Alternatively, Spearman rank correlation or mutual information could be used in order to examine the dependencies between time-series. In what follows we will discuss the details of our methodology and how the three general assumptions of causality studies (discussed in Section 2) are addressed.

**Conditional Ignorability Assumption:** We apply the Algorithm in order to find the set of time-series \( H \) that need to be controlled in order to satisfy the conditional ignorability assumption. According to our method, the resulted set contains all the direct predecessors of nodes \( X \) and \( Y \) that have in common. In figure 3, we depict the resulted set \( H \) of an example graph comprised by time-series \( X, Y \) and \( W \); the lags are as well as its lagged versions, with maximum lag \( L = 2 \). The parents of \( X \) and \( Y \) are selected by conducting conditional independence tests as described in Algorithm 1. For example, the arrow from \( X^{(1)} \) to \( X \) denotes that \( X \perp X^{(1)} \) and the arrow from \( X^{(2)} \) to \( X \) that \( X \perp X^{(2)}|X^{(1)} \). Similarly, the arrow from \( W \) to \( X \) denotes that \( X \perp W \) and the arrow from \( W^{(1)} \) to \( X \) denotes that \( X \perp W^{(1)}|W^{(1)} \). The lack of arrow from \( W^{(2)} \) to \( X \) denotes that \( X \perp W^{(2)}|W^{(1)} \). \( H \) includes all the common parents of \( X \) and \( Y \). Thus, all the variables that are correlated both with \( X \) and \( Y \) time-series are included; hence, the set \( H \) is sufficient. However \( H \) may include also redundant time-series, i.e., some of the time-series included in \( H \) may not correlate with \( X \) or \( Y \) conditional to a subset of \( H \). In causality studies based on regression, including redundant predictors on the model could result in overfitting and would jeopardize the validity of the conclusions. Moreover, the application of methods based on conditional independence tests using information theoretic approaches would be challenged by the inclusion of redundant covariates since it would require conditioning on large sets of variables. In contrast, studies based on matching are less affected by the inclusion of redundant confounding variables (spurious correlations). Several methods that enable matching on a large number of confounding variables have been proposed [10,12,13]. In addition, scientists are able to apply balance diagnostic tests in order to assess if any confounding bias has been adequately eliminated; consequently, false conclusions due to confounding bias can be diminished. Following the matching design, the set of time-series \( H \) is controlled by creating a set of pairs of time-samples \( G \) where each \( u \)-th time-sample with a positive treatment value \( X(t_u^x) \) is matched with a \( v \)-th time-sample with zero treatment value \( X(t_u^x) \) such that \( H(t_u^x) \approx H(t_v^x) \).

**Stable Unit Treatment Value Assumption:** Denote by \( P \) the set of time-series that are direct predecessors of the effect variable \( Y \). Assuming \( X \in P \) (if not, \( X \) is independent of \( Y \) and therefore there is no causation), the assumption is violated if \( X^{(l)} \in P \) and \( X^{(l)} \notin H \), for \( l > 0 \). Since units correspond to time-samples, \( X^{(l)} \in P \) implies that the outcome value \( Y(t_u^y) \) at time \( t_u^y \) depends on the value of the treatment time-series \( X \) at time \( t_u^{x,l} \). In order to satisfy the assumption, we modify the \( H \) set as follows:

\[ H := (X^{(1)},...,X^{(L)} \cap P) \cup H, \quad (2) \]

satisfying \( Y(t_u^y) \perp X(t_u^{x,l})|H(t_u^y), \forall u \neq v \).

i.i.d. **Assumption:** Denote by \( Y \) the value of the outcome variable for the time-samples that have a positive treatment value and with \( Y_0 \) for time-samples with zero treatment value. The average causal effect is estimated as \( \bar{E}(Y_1 - Y_0|H) \). In order to enable statistical inference, the variable \( \Delta Y := Y_1 - Y_0|H \) needs to be i.i.d.. If \( P \) the set of direct predecessors of \( Y \), the outcome value \( Y(t_u^y) \) of each time-
the following model:

$$t$$ respectively and with $$Z$$

$$u$$

$$X$$ efficient on avoiding false causal conclusions. We denote framework [43] and we demonstrate that our method is more
approach with a multivariate Granger causality model and

$$Y$$ modified set of time-series

$$H$$ := $$(Y^{(1)},...,Y^{(L)} \cap P) \cup H$$ (3)

Causal inference will be performed by matching on the
modified set of time-series $$H$$ thus, the variable $$\Delta Y := Y_1 - Y_0|H$$ will be i.i.d..

4. EVALUATION

4.1 Evaluation with Synthetic Data

In order to demonstrate the potential of our approach we assess its effectiveness in detecting causal relationships on linear and non-linear synthetic data. We also compare our approach with a multivariate Granger causality model and with an information theoretic approach based on Runge’s framework [43] and we demonstrate that our method is more efficient on avoiding false causal conclusions. We denote with $$X = \{X(t_u^x) : u = 1,2,...N\}$$ and $$Y = \{Y(t_u^y) : u = 1,2,...N\}$$ the treatment and outcome time-series respectively and with $$Z = \{Z(t_u^z) : u = 1,2,...N\}$$ a set of $$M$$ confounding variables. We also assume that $$t_u^x < t_u^y < t_u^z, \forall u$$. The relationships among $$X, Y$$ and $$Z$$ are described by the following model:

$$X(t^x_u) = h_{xz}(X(t^x_{u-1})) + f_{xz}(Z(t^z_u)) + \epsilon_x(t^x_u)$$ (4)

$$Y(t^y_u) = h_{yz}(Y(t^y_{u-1})) + f_{yz}(Z(t^z_u)) + f_{gy}(X(t^x_u)) + \epsilon_y(t^y_u)$$ (5)

$$Z^i(t^z_u) = h_{zi}(Z^i(t^z_{u-1})) + \epsilon_z(t^z_u), \forall Z^i \in Z, \quad (6)$$

where $$\epsilon_x(t^x_u), \epsilon_y(t^y_u)$$ and $$\epsilon_z(t^z_u)$$ are i.i.d. Gaussian noise variables with zero mean and std. dev. equal to $$20 + 2 \cdot M, 10 + 2 \cdot M$$ and 10, respectively.

We consider the following four cases:

**Case 1.** The model is linear. Thus, $$f_{xz}(Z(t^z_u)) = \sum_i \alpha_{xz,i} \cdot Z^i(t^z_u)$$, $$h_{xz}(X(t^x_{u-1})) = \alpha_{xz} \cdot X(t^x_{u-1})$$, $$h_{yz}(Y(t^y_{u-1})) = \alpha_{yz} \cdot Y(t^y_{u-1})$$.

**Case 2.** We apply the linear model of Case 1, but we set $$f_{yz}(X(t^x_u)) = 0$$. In this case the treatment time-series $$X$$ does not have any causal impact on the outcome time-series.

**Case 3.** The dependencies between the confounding variables and the treatment and effect variables are non-linear. In particular,

$$f_{xz}(Z(t^z_u)) = \sum_i \alpha_{xz,i} \cdot (Z^i(t^z_u))^2$$

$$f_{yz}(X(t^x_u)) = \sum_i \alpha_{yz,i} \cdot (Z^i(t^z_u))^2$$

We use the linear equations of Case 1 for the rest of the functions.

**Case 4.** We use the non-linear model of Case 3, but we set $$f_{yz}(X(t^x_u)) = 0$$. In this case, the multivariate linear Granger causality approach may return positive causality result, even though the treatment time-series $$X(t)$$ does not have any causal impact on the outcome time-series.

A unit (i.e. time-sample) $$u$$ of the study is described by the set of time-series values: $$S(t_u) := (X(t^x_u), Y(t^y_u), Z(t^z_u), X^{(1)}(t^x_u), Y^{(1)}(t^y_u), Z^{(1)}(t^z_u))$$. We apply the following three methodologies on the synthetic data generated using the models above in order to assess the causal impact of variable $$X$$ on $$Y$$:

**Multivariate Granger Causality (MGC).** We apply stepwise regression in order to fit our data to the following model:

$$Y(t^y_u) = a_1 \cdot Y(t^y_{u-1}) + \sum_{l=0}^{(1)} b_l \cdot X(t^x_{u-l}) + \sum_{l=0}^{(1)} c_l \cdot Z(t^z_{u-1}) + \delta + \epsilon(t^y_u)$$ (7)

We conclude that $$X$$ causes $$Y$$ if $$X$$ or any lagged version of $$X$$ is included in the regression model.
Conditional Mutual Information Tests (CMI). Following Runge’s approach [43], a causal graph is created by performing conditional independence tests using conditional mutual information as described in Section 2.3.

Matching Design for Time-series (MDT). Following the proposed approach, we apply Algorithm 1 in order to find the set of variables $H$ that needs to be controlled in order to achieve conditional ignorability. The resulting graph is depicted at Figure 4. $H$ includes any $Z' \in Z$ that correlates both with $X$ and $Y$. Moreover, we satisfy the i.i.d assumption by including in $H$ the time-series $Y^{(1)}$. In order to create groups of treated and untreated units we first transform the time series $X$ into a binary stream $\tilde{X}$: $\tilde{X}(t^o) = 0$, if $X(t^o) < \mu_X$; $\tilde{X}(t^o) = 1$, otherwise, where $\mu_X$ is the mean of $X$ (i.e. the $u$-th time-sample corresponds to a treated unit if $X(t^o) > \mu_X$). Then, we create pairs of treated and untreated units (i.e. time-samples) by applying Genetic Matching algorithm [46]. Genetic matching is a multivariate matching method which applies an evolutionary search algorithm in order to find optimal matches which minimize a loss function. We use as a loss function the average standardized mean difference between the treated and untreated units (i.e. time-samples) by reflecting real Twitter data. Our study considers the daily closings of the stock market and people trust on the company are reflected on Twitter data. We gather this information from uniform distribution on $[\mu_A - 0.5 \sigma_A, \mu_A + 0.5 \sigma_A]$. In Case 1 and Case 3, $A_k$ denotes the number of times that a causal relationship from $X$ to $Y$ is successfully inferred (true positive) for the $k$th set of model coefficients and different noise realizations, while in Case 2 and 4 it denotes the number of times that a causal relationship is falsely inferred (false positive). In Figure 5 we present the mean value of $A_k$, $\mu_A$, along with the standard error of the mean. According to our results, the proposed causal inference technique reduces significantly the number of false positive causality conclusions while it is slightly less successful on detecting real causality for $M = 10$. Multivariate Granger causality achieves almost 100% accuracy on true causality detection both for the linear (Case 1) and non-linear (Case 3) cases. However, it performs poorly in terms of avoiding false positive conclusions. The performance of all the examined methods improves for larger $M$ values (apart from multivariate Granger causality on the linear cases). This is due to the fact that, by adding more variables on the set $Z$, the dependence of $X$ and $Y$ with each individual $Z' \in Z$ is weaker; consequently, although $M$ covariates are used to generate $X$ and $Y$ time-series, for large $M$ values, only a subset of them has significant effect on them. Thus, canceling out the effect of $Z$ is easier.

4.2 Causal Effect of Social Media on Stock Markets

4.2.1 Dataset Description

We apply our method in order to investigate whether information about specific companies and people reactions influence stock market prices. We gather this information by analyzing relevant tweets. Twitter enables us to capture people opinions about the target companies, their optimism/pessimism about stock market movements and their reaction to news such as quarterly results announcements or new product launches. Thus, factors related to the company performance and people trust on the company are reflected on Twitter data. Our study considers the daily closing prices of four big tech companies traded on NASDAQ market: Apple Inc., Microsoft, Amazon.com and Yahoo!
We estimate a daily sentiment index for each of these companies by analyzing the sentiment of related tweets. Tweets are gathered using the the names and the ticker symbols of the examined companies. In detail, we use the hashtags #apple and #AAPL for Apple Inc., #microsoft and #MSFT for Microsoft, #amazone and #AMZN for Amazon.com and #yahoo and #YHOO for Yahoo!

Our study is based on data gathered for four years, from January 2011 to December 2014. We examine whether the sentiment of tweets that are posted before stock market closing time influences the closing prices of the target stocks. In order to eliminate any confounding bias we need to control for factors that may affect both humans sentiment and the target stock prices. Potential influential factor on stocks daily closing prices are their opening prices and their performance during the previous days. Several works have also demonstrated that the performance of other big companies (either local or overseas companies) could influence some stocks (see for example [12, 13]). Foreign currency exchange rates may also cause money flows to overseas markets and consequently influence stocks prices. Finally, commodities prices could affect the earnings of companies and, therefore, their stocks prices. More specifically, our study involves the following time-series:

The response time-series \( Y \). The difference on the closing prices of the target stocks between two consecutive days. The \( u \)-th time-sample \( t^u \) of the time-series \( Y \) corresponds to the closing value of the \( u \)-th day minus the closing value of the previous day.

The treatment time-series \( X \). A daily sentiment index that is estimated using tweets related to the target stocks that are posted up to 24 hours before the closing time of the corresponding stock market. In order to assure that the values of the treatment variable are driven by information that has been available before the closing time of the target stocks, we omit from the study tweets posted up to one hour before the closing time. Thus, the sentiment index of day \( u \) is estimated using all the tweets posted from 4:00 p.m. (ET) time (i.e. the NASDAQ closing time) of day \( u - 1 \) to 3:00 p.m. (ET) time of day \( u \). Consequently, our treatment variable captures the people sentiment and reactions to news realizes at any time during the day, up to one hour before the stock market closing time. Tweets are filtered using the name of the company and the stock symbol as keywords.

The set of time-series \( Z \). We consider the following time-series which might influence our case study:

1. **The difference between the opening and closing prices of two consecutive days.** This time-series is an indicator of the activity of the target stocks at the start of the trading day.

2. **The stock market prices of several major companies around the world.** In our study we include all the components of the most important stock market indexes such as NASDAQ-100, Dow-30, Nikkei 225, DAX and FTSE. The study could be influenced only by factors that precede temporally both the treatment and effect variables. Thus, we use the difference between the opening and closing prices of two consecutive days for stocks that are traded in the USA exchange markets. The closing time of companies traded at the overseas markets precedes the closing time of the USA stock exchange market, thus the time-series for all the overseas companies stocks correspond to the difference on the closing prices between two consecutive days. Although the values of the treatment variable are driven by tweets that are posted both before and after the corresponding values of the time-series that we use to describe the performance of big companies, for convenience, we consider that the \( u \)-th time-sample \( t^u \) of the treatment time-series occurs one hour before the USA stock exchange market closing time at day \( u \). Thus, the time-sample \( t^u \) of any of the time-series that are used to describe the performance of either a USA-based company or an overseas company temporally precedes the \( u \)-th sample of the treatment time-series.

3. **The daily opening values of foreign currency exchange rates minus the previous day opening values.** We include the exchange rates between dollar and British pound, Euro, Australian dollar, Japanese Yen, Swiss Franc and Chinese Yen.

4. **The difference between the opening values of commodities for consecutive days.** We include the following commodities: gold, silver, copper, gas and oil.

### 4.2.2 Daily Sentiment Index Estimation

We classify each tweet as negative, neutral or positive using the SentiStrength classifier [19]. SentiStrength estimates the sentiment of a sentence using a list of terms where each term is assigned a weight indicating its positivity or negativity. We updated the list of terms in order to include terms that are commonly used in finance.

Sentiment extraction from text may be inaccurate. Although this issue has been disregarded in previous works [3, 9], here, in order to account for such inaccuracies on sentiment classification, we estimate a probability distribution function of the daily sentiment instead of a single metric. Let us define a set of three objects \( S = \{ \text{positive}, \text{neutral}, \text{negative} \} \). Each object \( i \in S \) denotes a classification category. Let us also define a random variable \( V_i \) as follows:

\[
V_i = \begin{cases} 
0 & \text{if a negative tweet is classified in class } i \\
1 & \text{if a neutral tweet is classified in class } i \\
2 & \text{if a positive tweet is classified in class } i 
\end{cases} 
\] (10)

We derive the probability distribution functions of each random variable \( V_i \), with \( i \in S \), based on the classification performance results. We evaluate the performance of the classifier by manually classifying 1200 randomly selected tweets (200 tweets for each one of the four examined companies). The probability distribution functions are presented in Table I.

Let us define with \( N_i \) the number of tweets posted within a day that are classified in category \( i \). We define a random

\[1\]Using the ticker symbol as hashtag for downloading relevant tweets could be problematic for companies with one letter symbol. In such cases, researchers have to identify first the hashtags or names used to identify these companies and search them instead.

\[2\]The words list along with the updated words weights can be found here: https://www.cs.bham.ac.uk/ tkt357/Publications.html
We create a binary treatment variable $X$ by applying thresholds on $P_{pos}(t_u)$. More specifically, a unit that describes the $u$-th day of the study is considered to be treated (i.e. $X(t_u^+)=1$) if $P_{pos}(t_u^+)$ ≥ $P_{thresh}$ and untreated (i.e. $X(t_u^+) = 0$) if $P_{pos}(t_u^-) < P_{thresh}$. We conduct our study for three different pairs of thresholds. In detail, we consider a pair of thresholds $T1$, where thresholds $P_{thresh}$ and $P_{thresh}$ are set to the 50th percentile of $X$, a pair of thresholds $T2$ where $P_{thresh}$ is set to the 60th percentile of $X$ and $P_{thresh}$ to the 40th percentile of $X$ and finally a pair $T3$ where $P_{thresh}$ and $P_{thresh}$ are set to the 70th and 30th percentiles respectively. By increasing the value of $P_{thresh}$ and decreasing the value of $P_{thresh}$ we eliminate from our study days in which the estimated tweets polarity is uncertain either due to measurement error or because the overall sentiment that is expressed during these days is considered to be neutral. Although discretization of a continuous variable results in information loss which may jeopardize, in some cases, the reliability of the causal inference, we enhance the validity of our conclusions by considering different threshold values.

We include in our study all the previously mentioned variables. We found that there is no autocorrelation in our time-series, thus, since there is no dependence of our time-series on their past values, we set the maximum lag $L$ equal to 1. For each of the four target stocks, we applied Algorithm 4 in order to find the set of time-series $H$ that needs to be controlled. We consider a correlation to be statistically significant if the corresponding $p$-value is smaller than 0.05. We used Spearman’s rank correlation in order to capture potentially non-linear relationships among the examined variables. We found that stock movements are significantly correlated with the sentiment of tweets posted within the same day. Our findings are in agreement with results of other studies [3, 7, 8]. We also found that stock prices are independent to past tweets sentiment conditional to more recent tweets. This indicates that any effect of tweets on stock prices is instant rather than long-term. Finally, according to our results, the daily movement of the traded assets for the target companies does not correlate with past days movements. This finding is consistent with the weak-form efficient market hypothesis according to which, it is not feasible to predict stock market movements by applying technical analysis. In Table 2 we present the correlation coefficient of the effect variable $Y$ with the treatment variable $X$ and the 1-lagged variables $X(t_u)$ for each one of the four examined companies. In Figure 6, we present the empirical probability distribution function of having a positive movement on the traded assets prices conditional to the sentiment of the tweets $P(Y(t_u^+) > 0|X(t_u^+))$. The probability distribution function is estimated using data collectively for the four examined companies. Our results indicate that the probability of having a positive movement on the stock market does not increase linearly with the daily tweets positivity index. Stock market movement is quite uncertain when the positivity index of the tweets ranges between 0.35 and 0.65, while the probability of having a positive movement is increasing for positivity index larger than 0.65. Moreover, we notice a relatively high probability of having a positive movement in days with sentiment positivity index lower than 0.1. Considering that daily tweets sentiment capture the current and past stock market trends, this could be attributed to the fact that investors may consider that it is a good time to invest money when assets prices are low; consequently, this could give lead to an increase of stock market prices.

Moreover, we find that both the effect and the treatment variables correlate with the most recent stock prices of several local and overseas companies. The daily movements of the target stocks correlate with US dollar exchange rates; however, currency exchange rates do not have any impact.
Table 2: Correlation of $Y$ with $X$, $X^{(1)}$ and $Y^{(1)}$.

|       | AAPL | MSFT | AMZN | YHOO |
|-------|------|------|------|------|
| $X$   | 0.393| 0.155| 0.237| 0.273|
| $X^{(1)}$ | 0.032| 0.036| 0.012| 0.046|
| $Y^{(1)}$ | 0.009| -0.003| -0.037| 0.031|

on the treatment variable. In Table 3 we present the number of variables from each category that will be included in the set $H$ for the four target companies and in Figure 7 we present the correlation coefficients of the treatment and effect time-series with all variables in set $H$. For all the examined stocks, the strongest confounder is their opening prices.

Table 3: Number of variables that are included in the set $H$ for each of the four examined companies.

| Category         | AAPL | MSFT | AMZN | YHOO |
|------------------|------|------|------|------|
| Nasdaq-100 Comp. | 6    | 21   | 33   | 7    |
| Nikkei Comp.     | 1    | 3    | 1    | 13   |
| DAX Comp.        | 18   | 2    | 7    | 10   |
| FTSE Comp.       | 10   | 3    | 12   | 26   |
| Dow-30 Comp.     | 7    | 3    | 9    | 2    |
| FOREX            | 0    | 0    | 0    | 0    |
| Commodities      | 0    | 0    | 0    | 0    |

In order to eliminate the effect of the confounding variables we need to match treated and control units with similar values on their set of confounding variables. We create opti-
The main limitation of all non-experimental causality studies is that they are based on the assumption that all confounding variables are known. However, in real scenarios there may be unmeasured factors which influence the assignment of units to treatments. In such cases, the conditional ignorability assumption is violated and consequently, any causal inference result may be biased. In our study, we include a large number of potentially influential factors such as the performance of other companies traded assets, commodities prices and currency exchange rates. However, there are other factors, such as inflation rates, political changes or economic policy changes which could influence both people sentiment, captured through Twitter, and traded assets prices. Although such factors may be reflected on the observed confounding variables (e.g. macroeconomic factors such as inflation rates would also affect the prices of other traded assets and consequently, the observed Twitter sentiment may be independent to inflation rates conditional to the performance of other assets included in the study), there may still be some bias due to unobserved factors.

A sensitivity analysis can be conducted in order to assess how the results of the study would be influenced in the presence of unmeasured confounding variables. We denote with \( \pi_t \) the probability that a unit \( t \) corresponding to the \( n \)-th day is assigned to a treatment (i.e. \( X(t_n) = 1 \)) and \( O_t = \pi_t / (1 - \pi_t) \) the odds of the unit to receive a treatment. Then, we denote with \( \Gamma = O_t / O_{t_n} \) the ratio of the odds of two units \( t_n, t_u \). If \( \Gamma = n \), the unit \( t_n \) is \( n \) times more likely to receive a treatment than unit \( t_u \) due to unobserved factors. Under the conditional ignorability assumption (i.e. units are equally likely to receive a treatment conditional to their observed characteristics), the ratio \( \Gamma \) should be equal to 1 for two matched time-samples \( t_n \) and \( t_u \).

In [50], Rosenbaum applies the Wilcoxon’s signed rank test for the resulted matched treated and control pairs of a causality study under the null hypothesis that the treatment has no effect on the observed outcome variable. According to this method, for each matched pair \( (t_n, t_u) \) a rank is assigned to the outcome difference \( Y(t_n) - Y(t_u) \). The Wilcoxon’s signed rank statistic \( W \) is estimated as the sum of the ranks of the positive differences (the interested reader can find a detail description of the method in [51]). Under the null hypothesis, the mean value of \( W \) is \( S^2 \cdot (S + 1)/4 \), where \( S \) the number of matched samples. When \( S \) is sufficiently large, the upper bound of the distribution of \( W \) can be approximated by a normal distribution with mean \( (1 + \Gamma) \cdot S^2 \cdot (S + 1)/2 \). Thus, the sensitivity on unobserved confounding variables can be assessed by computing the upper bounds on the p-values of the Wilcoxon’s signed rank test for increasing \( \Gamma \) values.

We apply this method in order to evaluate the sensitivity of our results on unobserved confounding variables. At Table 5, we present the results of our sensitivity analysis for

| Company | MGC | CMI |
|---------|-----|-----|
| AAPL    | 0.0000 | 0.0002 |
| MSFT    | 0.0008 | 0.0021 |
| AMZN    | 0.0000 | 0.0001 |
| YHOO    | 0.0011 | 0.0004 |
Γ ≤ 2.0 and for the T2 pair of thresholds. According to our results, the causal influence of Twitter on Apple stock prices would be considered statistically significant (with p-value 0.014) even if some days were twice more likely (i.e. Γ = 2) to have positive sentiment conditional to the observed confounding variables due to unmeasured factors. Similarly, for Amazon our causal inference results are statistically significant (i.e. p-value < 0.05) for Γ ≤ 1.9, for Yahoo! for Γ ≤ 1.7 and finally for Microsoft our conclusion would be invalid for Γ ≥ 1.6.

Several works have previously examined the potential of information extracted from social media, search engine query data or other web-related information to predict stock market returns. For example, in [8] the authors demonstrate that the level of optimism/pessimism, which is estimated using Twitter data, correlates with stock market movement. Other projects have been focused on the possible use of sentiment analysis based on Twitter data for the prediction of traded assets prices by applying a bivariate Granger causality analysis [3, 9] or regression models [52]. Similarly, in [53] information theoretic methods are used to investigate whether sentiment analysis of social media can provide statistically significant information for the prediction of stock markets and in [54] authors propose a prediction method based on machine learning. In [55] the authors have also demonstrated that search engine query data correlate with stock market movements. In [56] the authors propose a trading strategy that utilizes information about Wikipedia views. They demonstrate that their trading strategy outperforms random strategy. However, all the above mentioned studies are based on bivariate models. Although their results indicate that social media and other web sources may carry useful information for stock market prediction, by using these techniques it is not possible to figure out if other factors are influencing the observed trends.

Trading strategies that utilize both technical analysis and sentiment analysis are discussed in [57, 58]. However, these works are based on regression analysis, thus they suffer from the limitations that have been previously discussed. Moreover, all the studies so far focus mainly on prediction of stock market movement. Although they provide insights about the influence that emotional and social factors may have on stock market, they do not investigate the presence of causality. To the best of our knowledge, this is the first work

\begin{table}[h]
\begin{center}
\begin{tabular}{|c|cccc|}
\hline
\multicolumn{1}{|c|}{\textbf{Γ}} & \textbf{AAPL} & \textbf{AMZN} & \textbf{YHOO} & \textbf{MSFT} \\
\hline
1.0 & 0.0000 & 0.0000 & 0.0000 & 0.0000 \\
1.1 & 0.0000 & 0.0000 & 0.0000 & 0.0000 \\
1.2 & 0.0000 & 0.0000 & 0.0000 & 0.0005 \\
1.3 & 0.0000 & 0.0002 & 0.0000 & 0.0005 \\
1.4 & 0.0001 & 0.0011 & 0.0003 & 0.0109 \\
1.5 & 0.0003 & 0.0042 & 0.0010 & 0.0327 \\
1.6 & 0.0009 & 0.0131 & 0.0034 & 0.0775 \\
1.7 & 0.0021 & 0.0328 & 0.0091 & 0.1519 \\
1.8 & 0.0043 & 0.0692 & 0.0209 & 0.2552 \\
1.9 & 0.0082 & 0.1263 & 0.0419 & 0.3789 \\
2.0 & 0.0142 & 0.2050 & 0.0749 & 0.5093 \\
\hline
\end{tabular}
\end{center}
\caption{Sensitivity Analysis.}
\end{table}
that attempts to measure the causal effect of such factors on stock markets.

6. DISCUSSION

Our results on the simulated experiments indicate that our method is more effective on avoiding false positive causality conclusions. We have examined the performance of the proposed method in datasets with up to 50 dimensions and 100 time-samples. Extreme high-dimensional cases with \( p > n \) are not considered in this study. In such cases, balancing treatment and control groups for each confounding variable would require large number of samples. Propensity score matching \([26]\) represents an alternative matching method that can effectively handle large number of confounding variables by performing matching on a single balancing score, i.e. the propensity score. The propensity score corresponds to the probability of a unit to be assigned to a treatment and it is usually approximated by applying a logistic regression model of the treatment against the set of confounding variables. High-dimensional propensity score matching \([59]\) has also been proposed in order to handle extreme high-dimensional cases with \( p >> n \).

One of the main advantages of the proposed MDT method over multivariate Granger causality and CMI is that the design of the study is separated from the analysis. The values of the response time-series \( Y \) are not used during the matching process. The causal impact of a time-series \( X \) on \( Y \) is evaluated only after sufficient balance between the treated and untreated samples has been achieved. In contrast, in a regression-based analysis the response time-series \( Y \) is used in order to learn the coefficients of the predictor variables of the study. Many studies suggest that regression-based methods for causal inference are less reliable \([60]\). Moreover, the proposed method is non-parametric, while Granger causality is based on assumptions about the model class (i.e. linear/non-linear relationships). According to our results, linear Granger causality performs poorly when there are non-linear relationships among the examined time-series.

In addition, as it was previously discussed, MDT requires significantly fewer conditional independence tests and smaller conditioning sets. In detail, the maximum conditioning set of the proposed method is equal to the maximum lag \( L \), while the maximum conditioning set of CMI is \( M \cdot L \) (with \( M \) the number of confounding variables). Thus, MDT can handle more effectively datasets which include large number of confounding variables.

Moreover, the computational cost of creating the graph is significantly lower for MDT compared to CMI. Assuming discrete time-series with values in a set \( V \), the computational complexity of creating a graph by applying the proposed method is \( O(|V|^L \cdot M \cdot |N|) \), while the computational cost of CMI is \( O(|V|^{|H| L} \cdot M \cdot |N|) \), with \(|V|\) the size of set \( V \). However, causal inference with MDT requires an additional matching step, and consequently, its computational complexity largely depends on the matching method that is applied. If a simple nearest neighbor matching method is applied \([16]\), the cost of finding the best match of a single unit is \( O(|H| \cdot |N|) \), with \(|H|\) denoting the size of set \( H \), and the cost of matching all the units is \( O(|H| \cdot N^2) \). Thus, the overall computational cost of MDT is \( O(|V|^L \cdot N + |H| \cdot N^2) \). However, when more complex matching methods, such as Genetic matching \([46]\), are applied, the computational cost of MDT can be significantly larger. Genetic matching algorithm applies an evolutionary search method in order to find optimal weights for each covariate. In each algorithm iteration, a set of \( P \) weights for each one of the variables in set \( H \) is generated. \( P \) corresponds to the population size of the genetic algorithm. Then, nearest neighbor matching is applied on the weighted variables of \( H \) for each one of the \( P \) weights. A loss function is used to estimate the loss for each of the \( P \) resulted sets of matched pairs. If the loss is sufficiently small for any of the \( P \) weights, the method terminates; otherwise this process is repeated. A maximum number of iterations \( I \) can be used in order to set an upper bound on the computational time of the method. In our experiments, we used as loss function the average standardized mean difference between the treated and control units. The cost of loss estimation for each weights set is \( O(|H| \cdot |N|) \). Thus, the total computational cost of the matching process is \( O(I \cdot P \cdot |H| \cdot N^2) \). Although the computational cost of MDT could be significantly larger than the cost of CMI, given the availability of advanced computational resources, the computational efficiency can be traded for more reliable results. In our simulated experiments, the running time of CMI was in order of seconds while the running time of MDT was in order of minutes, using a 2.6 GHz quad core CPU and 16 GB RAM.

Finally, we discuss the assumptions behind our method, thus outlining situations where the method is expected to perform well. There are four key ingredients in our method:

1. We perform independence tests on time series pairs. There is no guarantee that if there were higher-order dependencies among several time-series (e.g. the outcome variable and two confounding variables), they would be detected by the pair-wise tests.

2. In our conditional independence testing, the maximum conditioning set is determined by the largest lag \( L \). The value of \( L \) is of course upper bounded by our desire to have sample sizes large enough to yield sufficient power to independence tests.

3. The matching procedure assumes that there is an overlap in the confounding variables’ values between the groups of treated and control units. If this is not the case, the matching will not achieve sufficient balance. For example, in our case, it was not feasible to conduct a causality study of the impact of social media sentiment on the treated assets prices of Apple Inc. when the T3 thresholds pair is used, since there was not sufficient overlap on the confounding variables’ values.

4. The estimation of the average treatment effect is influenced by the power of the statistical test that is applied.

7. CONCLUSIONS

In this study, for the first time we have attempted to quantify the causal impact of social and emotional factors, captured by social media, on daily stock market returns of individual companies (i.e., not just a mere correlation between the two). We have proposed a novel non-parametric framework for causal analysis in time-series. Our evaluation on synthetic data demonstrates that our method is more effective on inferring true causality and avoiding false positive conclusions compared to other methods that have been
previously used for causal inference in time-series. Our approach can incorporate a large number of factors and, therefore, can effectively handle complex data such as financial data. Indeed, causality studies that are based on observational data rather than experimental procedures could be biased in case of missing confounding variables. However, conducting experimental studies is not feasible in most cases. In this work we have minimized the risk of biased conclusions due to unmeasured confounding variables by including in our study a large number of factors. Additionally, we conduct an analysis on the sensitivity of our conclusions on missing confounding variables. We have estimated a sentiment index indicating the probability that the general sentiment of a day, based on tweets posted for a target company, is positive. Our results show that Twitter data polarity does indeed have a causal impact on the stock market prices of the examined companies. Although our study involves only big technological companies and consequently our findings might not be generalizable, especially for companies that do not focus directly on retail customers. The analysis of the validity of these findings in the B2B sector is an open question that we plan to explore. Hence, we believe social media data could represent a valuable source of information for understanding the dynamics of stock market movements.

8. REFERENCES

[1] Sitaram Asur and Bernardo A Huberman. Predicting the future with social media. In Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT’10), pages 492–499. IEEE, 2010.

[2] Andranik Tumasjan, Timm Oliver Sprenger, Philipp G Sandner, and Isabell M Welpe. Predicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment. In Proceedings of the 4th International Conference on Weblogs and Social Media (ICWSM’10), volume 10, pages 178–185, 2010.

[3] Johan Bollen, Huina Mao, and Xiaojun Zeng. Twitter mood predicts the stock market. Journal of Computational Science, 2(1):1–8, 2011.

[4] R. M. Bond et al. A 61-million-person experiment in social influence and political mobilization. Nature, 489(7415):295–298, 2012.

[5] L. Muchnik, S. Aral, and S.J. Taylor. Social influence bias: A randomized experiment. Science, 341(6146):647–651, 2013.

[6] John Concato, Nirav Shah, and Ralph I Horwitz. Randomized, controlled trials, observational studies, and the hierarchy of research designs. New England Journal of Medicine, 342(25):1887–1892, 2000.

[7] Tobias Preis, Daniel Reith, and H Eugene Stanley. Complex dynamics of our economic life on different scales: insights from search engine query data. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 368(1933):5707–5719, 2010.

[8] Xue Zhang, Hauke Fuehres, and Peter A Gloor. Predicting stock market indicators through Twitter "I hope it is not as bad as I fear". In Proceedings of the 2nd Collaborative Innovation Networks Conference, volume 26, pages 55–62. Elsevier, 2011.

[9] Xue Zhang, Hauke Fuehres, and Peter A Gloor. Predicting asset value through Twitter buzz. In Advances in Collective Intelligence, pages 23–34. Springer, 2012.

[10] Clive WJ Granger. Some recent development in a concept of causality. Journal of Econometrics, 39(1):199–211, 1988.

[11] Huina Mao, Scott Counts, and Johan Bollen. Predicting financial markets: Comparing survey, news, twitter and search engine data. arXiv preprint arXiv:1112.1051, 2011.

[12] Dror Y Kenett, Michele Tumminello, Asaf Madi, Girit Gur-Gershgoren, Rosario N Mantegna, and Eshel Ben-Jacob. Dominating clasp of the financial sector revealed by partial correlation analysis of the stock market. PloS one, 5(12):e15032, 2010.

[13] K Tse Chi, Jing Liu, and Francis CM Lau. A network perspective of the stock market. Journal of Empirical Finance, 17(4):659–667, 2010.

[14] Jianqiong Fan, Fang Han, and Han Liu. Challenges of big data analysis. National Science Review, 1(2):293–314, 2014.

[15] William R. Shadish, Thomas D Cook, and Donald Thomas Campbell. Experimental and quasi-experimental designs for generalized causal inference. Wadsworth Cengage learning, 2002.

[16] E. A. Stuart. Matching methods for causal inference: A review and a look forward. Statistical Science: a Review Journal of the Institute of Mathematical Statistics, 25(1):1, 2010.

[17] P. W. Holland. Statistics and causal inference. Journal of the American Statistical Association, 81(396):945–960, 1986.

[18] Stephen L Morgan and Christopher Winship. Counterfactuals and causal inference. Cambridge University Press, 2014.

[19] Peter M Bentler. Multivariate analysis with latent variables: Causal modeling. Annual review of psychology, 31(1):419–456, 1980.

[20] Paul R Rosenbaum and Donald B Rubin. Assessing sensitivity to an unobserved binary covariate in an observational study with binary outcome. Journal of the Royal Statistical Society. Series B (Methodological), pages 212–218, 1983.

[21] David Freedman. From association to causation via regression. Advances in Applied Mathematics, 18(1):59–110, 1997.

[22] Guido W Imbens. Nonparametric estimation of average treatment effects under exogeneity: A review. Review of Economics and Statistics, 86(1):4–29, 2004.

[23] Peter C Austin. The relative ability of different propensity score methods to balance measured covariates between treated and untreated subjects in observational studies. Medical Decision Making, 29(6):661–677, 2009.

[24] Peter C Austin. Goodness-of-fit diagnostics for the propensity score model when estimating treatment effects using covariate adjustment with the propensity score. Pharmacoepidemiology and Drug Safety, 17(12):1202–1217, 2008.

[25] Peter C Austin. Assessing balance in measured baseline covariates when using many-to-one matching
on the propensity-score. *Pharmacoepidemiology and Drug Safety*, 17(12):1218–1225, 2008.

[26] Valerie S Harder, Elizabeth A Stuart, and James C Anthony. Propensity score techniques and the assessment of measured covariate balance to test causal associations in psychological research. *Psychological methods*, 15(3):234, 2010.

[27] Judea Pearl. An introduction to causal inference. *The International Journal of Biostatistics*, 6(2), 2010.

[28] Judea Pearl. *Causality: models, reasoning and inference*. Cambridge University Press, 2009.

[29] Judea Pearl. Causal diagrams for empirical research. *Biometrika*, 82(4):669–688, 1995.

[30] Markus Kalisch and Peter Bühlmann. Estimating high-dimensional directed acyclic graphs with the pc-algorithm. *Journal of Machine Learning Research*, 8(Mar):613–636, 2007.

[31] Junning Li and Z Jane Wang. Controlling the false discovery rate of the association/causality structure learned with the pc algorithm. *Journal of Machine Learning Research*, 10(Febr):475–514, 2009.

[32] Peter Spirtes. An anytime algorithm for causal inference. In *AISTATS*. Citeseer, 2001.

[33] Diego Colombo, Marloes H Maathuis, Markus Kalisch, and Thomas S Richardson. Learning high-dimensional directed acyclic graphs with latent and selection variables. *The Annals of Statistics*, pages 294–321, 2012.

[34] Clive WJ Granger. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica: Journal of the Econometric Society*, pages 424–438, 1969.

[35] Adam B Barrett, Lionel Barnett, and Anil K Seth. Multivariate granger causality and generalized variance. *Physical Review E*, 81(4):041907, 2010.

[36] Kun Zhang, Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. Kernel-based conditional independence test and application in causal discovery. arXiv preprint arXiv:1202.3773, 2012.

[37] Jakob Nawrath, M Carmen Romano, Marco Thiel, István Z Kiss, Mahesh Wickramasinghe, Jens Timmer, Jürgen Kurths, and Björn Schelter. Distinguishing direct from indirect interactions in oscillatory networks with multiple time scales. *Physical Review Letters*, 104(3):038701, 2010.

[38] Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. Causal inference on time series using restricted structural equation models. In *Advances in Neural Information Processing Systems*, pages 154–162, 2013.

[39] Doris Entner and Patrik O Hoyer. On causal discovery from time series data using fci. *Probabilistic graphical models*, pages 121–128, 2010.

[40] Thomas Schreiber. Measuring information transfer. *Physical Review Letters*, 85(2):461, 2000.

[41] Lionel Barnett, Adam B Barrett, and Anil K Seth. Granger causality and transfer entropy are equivalent for gaussian variables. *Physical review letters*, 103(23):238701, 2009.

[42] Bernd Pompe and Jakob Runge. Momentary information transfer as a coupling measure of time series. *Physical Review E*, 83(5):051122, 2011.

[43] Jakob Runge, Jobst Heitzig, Vladimir Petoukhov, and Jürgen Kurths. Escaping the curse of dimensionality in estimating multivariate transfer entropy. *Physical review letters*, 108(25):258701, 2012.

[44] Bo Lu, Elaine Zanutto, Robert Hornik, and Paul R Rosenbaum. Matching with doses in an observational study of a media campaign against drug abuse. *Journal of the American Statistical Association*, 96(456):1245–1253, 2001.

[45] Keisuke Hirano and Guido W Imbens. The propensity score with continuous treatments. *Applied Bayesian modeling and causal inference from incomplete-data perspectives*, 226164:73–84, 2004.

[46] Alexis Diamond and Jasjeet S Sekhon. Genetic matching for estimating causal effects: A general multivariate matching method for achieving balance in observational studies. *Review of Economics and Statistics*, 95(3):932–945, 2013.

[47] Jasjeet S Sekhon. Opiates for the matches: Matching methods for causal inference. *Annual Review of Political Science*, 12:487–508, 2009.

[48] Donald B Rubin. Estimating causal effects from large data sets using propensity scores. *Annals of Internal Medicine*, 127(8_Part_2):757–763, 1997.

[49] Mike Thelwall. Heart and soul: Sentiment strength detection in the social web with sentistrength. *Cyberemotions*, pages 1–14, 2013.

[50] Paul R Rosenbaum. Observational studies. In *Observational Studies*, pages 1–17. Springer, 2002.

[51] Erich Leo Lehmann and HJ D’abreka. *Nonparametrics: statistical methods based on ranks*. Holden-Day, 1975.

[52] Jianfeng Si, Arjun Mukherjee, Bing Liu, Qing Li, Huayi Li, and Xiaotie Deng. Exploiting topic based twitter sentiment for stock prediction. *ACL (2)*, 2013:24–29, 2013.

[53] Ilya Zheludev, Robert Smith, and Tomaso Aste. When can social media lead financial markets? *Scientific Reports*, 4, 2014.

[54] Alexander Porshnev, Ilya Redkin, and Alexey Shevchenko. Machine learning in prediction of stock market indicators based on historical data and data from twitter sentiment analysis. In *2013 IEEE 13th International Conference on Data Mining Workshops*, pages 440–444. IEEE, 2013.

[55] Tobias Preis, Helen Susannah Moat, and H Eugene Stanley. Quantifying trading behavior in financial markets using Google Trends. *Scientific Reports*, 3, 2013.

[56] Helen Susannah Moat, Chester Curme, Adam Avakian, Dror Y Kenett, H. Eugene Stanley, and Tobias Preis. Quantifying wikipedia usage patterns before stock market moves. *Scientific Reports*, 3, 2013.

[57] Robert P Schumaker and Hsinchun Chen. Textual analysis of stock market prediction using breaking financial news: The azfin text system. *ACM Transactions on Information Systems (TOIS)*, 27(2):12, 2009.
price prediction. In Proceedings of IEEE Ninth International Conference on Dependable, Autonomic and Secure Computing (DASC’11), pages 800–807. IEEE, 2011.

[59] Sebastian Schneeweiss, Jeremy A Rassen, Robert J Glynn, Jerry Avorn, Helen Mogun, and M Alan Brookhart. High-dimensional propensity score adjustment in studies of treatment effects using health care claims data. Epidemiology (Cambridge, Mass.), 20(4):512, 2009.

[60] Peter C Austin. An introduction to propensity score methods for reducing the effects of confounding in observational studies. Multivariate Behavioral Research, 46(3):399–424, 2011.