



University of St.Gallen

SCHOOL OF MANAGEMENT, ECONOMICS, LAW, SOCIAL SCIENCES AND  
INTERNATIONAL AFFAIRS

# The effect of Twitter activity on Bitcoin price fluctuation

SOFTWARE ENGINEERING FOR ECONOMISTS  
(7,610,1.00)

*Alen Stepic - 11-475-258*  
*Dimitrios Koumnakes - 10-613-370*  
*Joël Sonderegger - 11-495-488*  
*Severin Kranz - 13-606-355*  
*Chi Xu - XX-XXX-XXX*

Fall Term 2017

Supervisor:  
Prof. Dr. Philipp Zahn  
FGN HSG  
Varnbühlstrasse 19  
9000 St. Gallen

29th December 2017

## **Abstract**

This paper, examines the dynamics between the amount of twitter activity regarding bitcoin and the actual price fluctuation of bitcoin. For the analysis two sets of data have been collected. One reflecting twitter activity and the other one showing the bitcoin prices for the same period. Based on that an econometric model is used for the statistical analysis and the interpretation of the results.

INPUT ALEN @DIMI: ADD JUST KEY FINDING THE REST I LIKE

This work was created in the context of a programming course for economists at the university of St. Gallen. To not exceed the framework of this study we require extensive knowledge in econometrics and take concepts as known. The underlying goal was to get familiar with software engineering tools and project management. Therefore, the academic content of this work is neither completed nor concluding.

# Contents

<b>1</b>	<b>Introduction</b>	<b>1</b>
1.1	Research Question and Goal of the Paper . . . . .	1
1.2	Methodology . . . . .	1
1.3	Scope . . . . .	2
<b>2</b>	<b>Brief Background and Data Collection</b>	<b>3</b>
<b>3</b>	<b>Econometric Modelling and Results</b>	<b>5</b>
<b>4</b>	<b>Conclusion</b>	<b>6</b>
<b>5</b>	<b>References</b>	<b>7</b>
<b>6</b>	<b>Declaration of Authorship</b>	<b>8</b>

## List of Figures

## List of Abbreviation

<b>VAR</b>	Vector Autoregression
<b>EMH</b>	Efficient Market Hypothesis
<b>etc</b>	et cetera

# 1 Introduction

Malkiel and Fama (1970, p. 388) introduced the Efficient Market Hypothesis (EMH), where they claim that under certain market conditions actual prices include all information. This hypothesis is broadly accepted in the financial world. The digitalisation leads to an increased network effect, where information can be share within second over the globe. Based on Mao, Counts and Bollen (2015, p. 3, as cited in Shiller, 2015, pp. TBD; Kahneman and Tversky, 2013, pp. TBD) the EMH fails to address the behavioural and emotional role of investors . The large fluctuation of the bitcoin prices in the last year (mostly increasing) and the fact that Twitter has become a very popular social network, where people interact, led to different research in the field of sentimental analysis with the focus on bitcoin and twitter. Mao et al. (2015, p. 18), claim that sentiment analysis using twitter tweets which contain the buzzword bullishness can indeed be used as an sentiment indicator for stock prices. As digitalization is expected to continue in the future, cryptocurrencies will hypothetical get more important. Thus, the topic is characterized as relevant.

## 1.1 Research Question and Goal of the Paper

As existing research focus mostly on the sentimental analyses of tweet content, this paper aims to examine existence of a correlation between the price development of bitcoin and the amount people talk about bitcoin on twitter. The goal should be reached by answering the following research question:

In what extend does the amount of twitter tweets influence the price of bitcoin?

## 1.2 Methodology

To answer the above mentioned research question a scientific approach has been applied. The paper consists out of three parts. The first part contains an introduction into the topic by pointing out the relevance of the topic, the research question, methodology and the scope. This is conducted in chapter 1. The second part is characterized as the theory part. The theory aims to provide to the reader the necessary background information by an introduction of relevant bitcoin and twitter information in context of the paper. The conducted research is based on forward research and using relevant sources. Furthermore, it contains a short introduction into the Vector Autoregression approach (VAR) and the constrains in context with the paper. This is conducted in chapters 2 and 3. The third part contains the discussion of the results and a conclusion. The conclusion is based on the theory and the own conducted mathematical computation. Those computations are based on a data set, which was gathered by own coded python scripts for the timeperiod December 20th - 27th. Based on the short time of observation the results have to be interpreted carefully. The data set contains out of two data sources (1) twitter tweets which contain the buzzword "bitcoin" and (2) historical bitcoin prices. Those two sources has been aggregated with a separate python script. The output of aggregation was

further proceeded with stata to conduct the VAR on a daily basis. The third part of the paper is discussed in the chapters 3 and 4.

### **1.3 Scope**

The paper's focus lies on the scientific discussion and answer of the research question mentioned above. The research question has consciously been design very tiny, as the focus of the course Software Engineering for Economists is the application of different tools, the documentation and reproducibility of the results, rather than writing a high quality scientific paper. As already mentioned above, the results have to be interpreted carefully because of the limited observation time of seven days. The approach and technical issues (like code scripts etc.) are not addressed. Those points are explained in the separate documentation paper.

## 2 Brief Background and Data Collection

### *Bitcoin*

After the global financial crisis, Satoshi Nakamoto (2008, p. 1) claimed, that a purely peer-to-peer version of electronic cash could bypass financial institutions as third parties for commercial transactions. Starting from his whitepaper, Bitcoin – the first electronic payment system that relies on the cryptographic concatenation of ongoing transactions (blockchain) has been developed (Nakamoto, 2008, p. 1).

Today, Bitcoin is the most popular of over 1300 so-called cryptocurrencies worldwide and increased in value over 1700 percent within the last year according to Coinmarketcap (<https://coinmarketcap.com/>). Even if the daily transaction volume increased rapidly over time (<https://coinmarketcap.com/>), this does not necessarily mean that Bitcoin is used for payment purposes. Even more, several authors argue that the Bitcoin price and transaction volumes are mainly pushed by speculative investments rather than actual usage as a currency (Corbet, Lucey & Yarovya, 2017; Forbes, 2017; Yermack, 2013). Kristoufek (2015) points out that there are other influencing factors such as usage in trade, money supply and price level, that influence the Bitcoin price in the long term. However, evidence shows that the level of Bitcoin prices are clearly driven by investors' interest in the digital currency. Kristoufek (2015) further explains the correlation is most evident in the long run. During explosive increases or rapid declines of the price higher investors' interest further boosts the movement in the direction (Kristoufek, 2015). These findings are in line with other researchers (see Garcia, Tessone, Mayrodiiev & Perony, 2014; Kondor, Posfai, Csabai, & Vattay, 2014).

One core assumption of behavioural finance is that investors' interests influence their behaviour and therefore stock prices. (Mao et al., 2015, p. 3) A widely used approach to measure investors' interests is the sentiment analysis of Twitter data. Based on analysis of the influence of twitter bullishness on the stock market by Mao et al. (2015), this paper further elaborates on the concept by applying the approach to Bitcoin prices.

### *Twitter*

(Text for Twitter) Input Alen: Explain why twitter is the right source and and some facts about it

### *Data Collection*

For our dataset, we used 2 different sources. We first collected historical data on bitcoin prices to USD and secondly, we gathered twitter messages related to bitcoin. We fulfil this relationship by selecting tweets that contain the term "bitcoin" in the text message. The time period for both collected data sets was from 21.12.2017 16:00:00 UTC to 26.12.2017 16:00:00 in an hourly basis. We chose an hourly basis to be able to do a statistical analysis with a certain amount of observation for the restricted processing time of this homework.

For the bitcoin prices, a publicly available API from CoinDesk was used (Source? See docu-

mentation?). The prices collected represent the final price level of bitcoin before every hour changes. For example, our first listed price of bitcoin was on 21.12.2017 at 16:59:59 and our last value on 26.12.2017 at 15:59:59. These values construct our first time series data set and are shown in the following graph.



### 3 Econometric Modelling and Results

As stated in the beginning, we want to examine the underlying relationship between twitter activity and bitcoin price fluctuations. From our data collection, we are left with two time series datasets representing two variables. A broadly used statistical method to simultaneously analyze multiple time series is the VAR (Vector Autoregression) approach. In this approach the endogenous variables are determined both by their own historical values and by the historical values of the other endogenous variables (Lütkepohl, 2005, p. 4-5).

To generate the econometric results that follow we used the statistical software STATA.

#### *Stationarity*

Before estimating a VAR model, the time series data must be checked for stationary. Thus, the means and variances are constant over time and the datasets do not show any trending behavior. Non-stationary data can lead to an inaccurate model which is undesirable. To test for stationarity, we use the Augmented Dickey-Fuller (ADF) test and get following results from STATA.

Interpreting the results and accepting or rejecting stationarity.

#### *Lag Specification*

To select the optimal number of lags in our VAR regression we check for various information criteria. Information criteria are measuring the tradeoff between model fit and parsimony, giving use the optimal number of lag to use (Brandt and Williams, 2007, p. 27). The calculations of these criteria for our given data set is easily done by a statistical software and given in the following table.

#### *VAR Regression Model*

In this paper, we use a basic unrestricted VAR model which consists of two endogenous variables, T for aggregate tweets and B for bitcoin prices. The selected time lag is 3 and this choice will be discussed later. Given that the model equations can be written as follows:

#### *Granger Causality*

To assess the causal relationship between our two endogenous variables and interpret the result we use the Granger causality. This test confirms if one variable is statistically useful to predict the other variable. If this is given the dynamics stated by the calculated coefficient above can be assumed to be valuable.

A second analysis to interpret our results is the impact response analysis. This evaluates the impact of changes in the one variable to the other variable. This is also provided by STATA and the results can be seen in the following graphs.

#### *Impulse Response Analysis*

## 4 Conclusion

(Insert text)

## 5 References

- Kahneman, D. & Tversky, A. (2013). Prospect theory: An analysis of decision under risk. In *Handbook of the fundamentals of financial decision making: Part i* (pp. 99–127). World Scientific.
- Malkiel, B. G. & Fama, E. F. (1970). Efficient capital markets: A review of theory and empirical work. *The journal of Finance*, 25(2), 383–417.
- Mao, H., Counts, S. & Bollen, J. (2015). *Quantifying the effects of online bullishness on international financial markets*. ECB Statistics Paper.
- Shiller, R. J. (2015). *Irrational exuberance*. Princeton university press.

## 6 Declaration of Authorship

We hereby declare,

- that we have written this thesis without any help from others and without the use of documents and aids other than those stated above;
- that we have mentioned all the sources used and that we have cited them correctly according to established academic citation rules;
- that we have acquired any immaterial rights to materials we may have used such as images or graphs, or that we have produced such materials ourself;
- that the topic or parts of it are not already the object of any work or examination of another course unless this has been explicitly agreed on with the faculty member in advance and is referred to in the thesis;
- that we will not pass on copies of this work to third parties or publish them without the University's written consent if a direct connection can be established with the University of St.Gallen or its faculty members;
- that we are aware that our work can be electronically checked for plagiarism and that we hereby grant the University of St.Gallen copyright in accordance with the Examination Regulations in so far as this is required for administrative action;
- that we are aware that the University will prosecute any infringement of this declaration of authorship and, in particular, the employment of a ghostwriter, and that any such infringement may result in disciplinary and criminal consequences which may result in our expulsion from the University or us being stripped of our degree.

.....  
Dimitrios Koumnakes - 10-613-370

.....  
Severin Kranz - 13-606-355

.....  
Joël Sonderegger - 11-495-488

.....  
Alen Stepic - 11-475-258

.....  
Chi Xu - XX-XXX-XXX

By submitting this academic term paper, we confirm through my conclusive action that we are submitting the Declaration of Authorship, that we have read and understood it, and that it is true.