



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 - 16-300-915*

Fall Term 2017

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

4th January 2018

## **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.

The evaluation of the model showed that changes in twitter activity on bitcoin have a minimal and insignificant effect on the future price development of bitcoin. Accordingly, aggregate twitter fees are not a good source to predict changes in bitcoin prices. This result must be understood with caution, because it only applies for the specific time interval of our observations, which is about 8 days and not for the past or future trend of cryptocurrencies.

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>Background and Data Collection</b>	<b>3</b>
2.1	Bitcoin . . . . .	3
2.2	Twitter . . . . .	4
2.3	Data Collection . . . . .	4
<b>3</b>	<b>Econometric Modelling and Results</b>	<b>6</b>
3.1	Stationarity . . . . .	6
3.2	Lag-Specification . . . . .	7
3.3	Regression Model . . . . .	8
3.4	Granger Causality . . . . .	9
<b>4</b>	<b>Conclusion</b>	<b>11</b>
<b>5</b>	<b>References</b>	<b>12</b>
<b>6</b>	<b>Declaration of Authorship</b>	<b>13</b>

## List of Figures

1	Twitter variables development . . . . .	5
2	Bitcoin price development . . . . .	5
3	Twitter variables stationarity test . . . . .	6
4	Bitcoin variables stationarity test . . . . .	7
5	Lag Specification . . . . .	8
6	Granger causality test . . . . .	9

## List of Tables

1	Var Regression . . . . .	9
---	--------------------------	---

## 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 shared over the globe. Based on Mao, Counts and Bollen (2015, p. 3, as cited in Shiller, 2015, pp. 175–195; Kahneman and Tversky, 2013, pp. 279) 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 a sentiment indicator for stock prices. As modern economy becomes more digitized, cryptocurrencies as a financial asset draws more attention, our paper seeks to shed some light on whether the tweets can be used to predict bitcoin prices.

## 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 frequency of bitcoin topics on twitter. The goal should be reached by answering the following research question:

In what extend does the amount of twitter tweets about bitcoin has an effect on the price movement 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 the reader with the necessary background information by introducing 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 (VAR) approach and its constraints 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 our mathematical computation. The computation is based on a data set, which was gathered by coded python scripts for the time period December 20th - 27th. Since the results are based on the observation of a short period, they have to be interpreted carefully. The data set consists of two data sources with

the following characteristics: (1) twitter tweets which contain the buzzword "bitcoin" and (2) those containing historical bitcoin prices. The two sources has been aggregated with another python script. The output of aggregation was further processed 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 of empirical test results. The research question has consciously been narrowed down, as the focus of the course Software Engineering for Economists is the application of different tools, the documentation and reproducibility of the results. As already mentioned above, the results have to be interpreted carefully because of the limited observation period of seven days. However, the results provide a first impression and can be further be developed. The approach and technical issues (like code scripts etc.) are not addressed, for purposes we shall address in a separate documentation paper.

## 2 Background and Data Collection

### 2.1 Bitcoin

In the aftermath of 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 this whitepaper, Bitcoin – the first electronic payment system that relies on the cryptographic concatenation of ongoing transactions<sup>1</sup> has been developed (Nakamoto, 2008, p. 1).

Today, Bitcoin is the most popular of thousands of 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 a means of payment. Even more, several authors argue that the Bitcoin price and transaction volumes are mainly driven by speculative investments rather than its actual use as a currency (Corbet, Lucey & Yarovya, 2017; Forbes, 2017; Yermack, 2013). Kristoufek (2015) points out that there are other influencing factors such as its use in trade, the supply of currencies and price level, which all contribute to determining the Bitcoin price in the long term. However, according to the same author, evidence shows that investors' interest plays an even greater role in bitcoin's price movements, and their correlation is even more evident in the long run, and during the periods of explosive increases or rapid declines of the price, investors' interest accelerates the price movements (Kristoufek, 2015). These findings are in line with other research findings (see Garcia, Tessone, Mayrodiiev & Perony, 2014; Kondor, Posfai, Csabai, & Vattay, 2014).

In previous research Madan, Saluja, and Zhao (2015) applied *machine learning* algorithms to predict Bitcoin prices. By using the *Bayesian regression*, Shah, Devavrat and Zhang (2015) achieved a return on investment of 89 percent over an investment period of 50 days. However, research in this area is still limited and most of them do not take into consideration the individuals' sentiments (Colianni, Rosales & Signorotti, 2015, p. 1).

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 done through sentiment analysis of Twitter data, which will be covered in the following section.

---

<sup>1</sup>blockchain

## 2.2 Twitter

With the current growth of social networks, the number of global social media users reached 2.46 billion and is expected to increase to 3.02 billion by 2021 according to Statista (2018). Gabriel and Röhrs (2017) define social networks as a loose connection of people in an online or internet community, respectively in a computer-based communication network (p. 12). This paper exploits the possibility to analyse shared “thoughts, views and opinions” (Stenqvist and Lönnö, 2017, p. 6) provided by such communities or communication networks.

Twitter was founded in 2006 and gained rapidly worldwide popularity (Twitter 2018). and reached 330 million active users in the third quarter of 2017 (Twitter 2017, Statista 2018). As a micro-blogging platform, Twitter suits the purpose of analysing investors’ interests due to its special characteristics. In comparison to other social networks, twitter limits its posts (tweets) to relatively short messages of 140 characters. Tweets can contain observations, thoughts, links to interesting content, websites, uploaded pictures or videos. Special conventions such as hashtags “#” or mentions “@” are used to reference searchable content or user profiles.(Schmidt, 2017)

In this way, Twitter users create millions of posts that contain interests, opinions and informations in both a private and professional context. Due to the semi-structured form, the message length restriction and classifying nature, Stenqvist and Lönnö (2017, p. 6) argue, that “Twitter has become a gold mine for opinionated data”, which is further supported in the wide adoption of researchers to analyse sentiments.

In relation to Bitcoin, researchers already showed that analysing investors sentiments can lead to reasonable correlations between Twitter sentiments and prices of cryptocurrencies (Stenqvist and Lönnö, 2017, p. 7), by either using a *polarity classification* – classifying the language in tweets as either positive or negative (see Colianni et al., 2015) – or a *lexicon based approach* – attributes predefined words to specific sentiment classifications (see Stenqvist and Lönnö, 2017). Hence, this paper choses a different focus area by limiting its scope to the number of daily tweets that contain the term “Bitcoin” without further analysing investors interests implied in the semantics. Inspired by analysis of the influence of twitter bullishness on the stock market by Mao et al. (2015), this paper applies the similar method to Bitcoin prices.

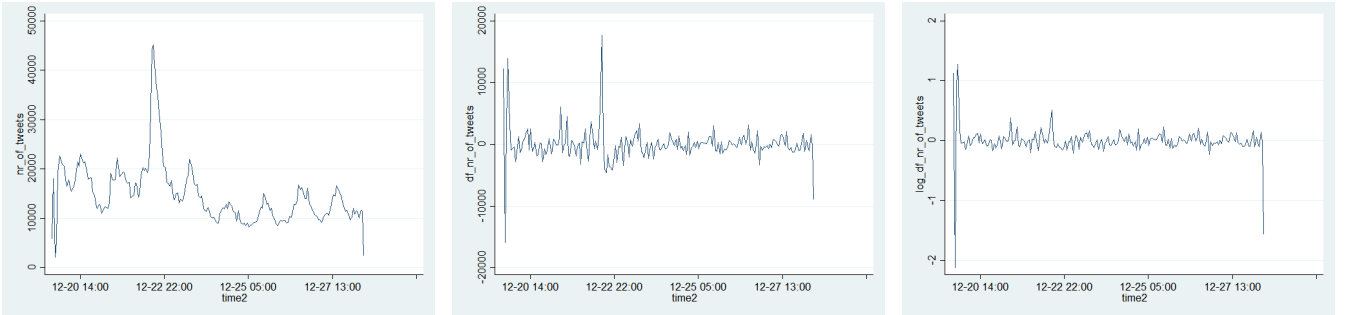
## 2.3 Data Collection

For our data, we collected 2 separate time series datasets. One showing the twitter activity regarding bitcoin, with the second one showing the historic bitcoin price development. Both datasets were collected hourly for the period from the 19.12.2017 starting at 20:00:00 UTC and



ending on the 28.12.2017 at 10:00:00 UTC. We chose the hourly basis to conduct statistical analysis with a limited amount of observations as demanded by the feasibility of such a complex project. For the twitter data we aggregated by each hour all tweets mentioning "bitcoin". We used prices at the last second of every hour to represent the hourly bitcoin price.

The following figure displays the twitter activity with regard to bitcoin for the observed time period. The first plot shows the hourly change of the aggregate number of tweets containing "bitcoin". In the second and third plots we respectively did a first difference and a logarithmic first difference transformation of our main variable, in case it is needed in the later analysis. The graphs show overall a cyclical movement around 1200 tweets and a single outburst befalling on 22.12.2017 between 15:00:01 and 21:00:00, with highest value being 4622 tweets in one hour.



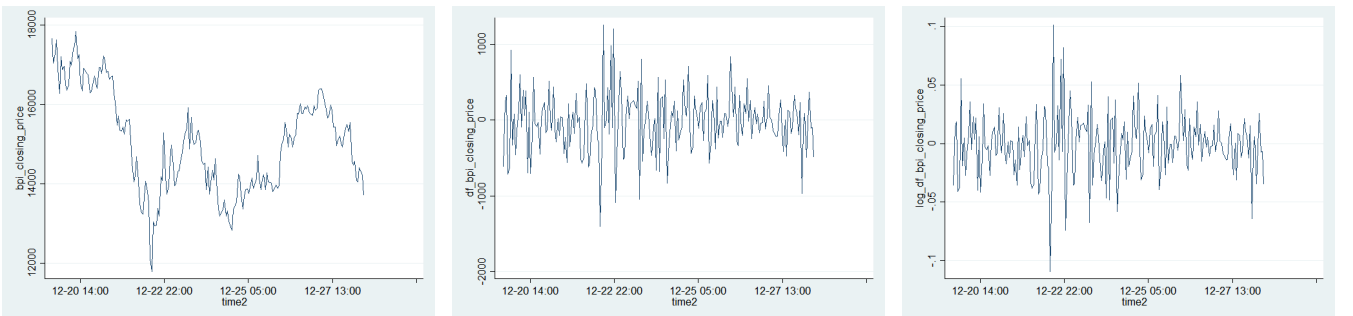
(a) Number of Tweets regarding bitcoin

(b) First difference of number of Tweets

(c) Log first difference of number of Tweets

Figure 1: Twitter variables development

Next we display the time series data on bitcoin price development for the observed time period. The first plot shows the hourly changes of bitcoin prices. The second and third plots are once again first difference transformations of the main variable for later analysis. In this graph we observe a notably volatile movement with no specific direction. This contradicts the general rising trend of bitcoin prices of the past months, but the problem may indeed arise if we happen to select a short period of one week when bitcoin price movement behaves less bullishly as usual, in other words, the lack of articulated trend may be attributed to *selection bias*.



(a) Hourly development of bitcoin prices

(b) First difference of bitcoin prices

(c) Log first difference of bitcoin prices

Figure 2: Bitcoin price development

### 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 Vector Autoregression (VAR) approach. In this method the endogenous variables are determined both by their own historical values and by the historical values of the other endogenous variables (Lütkepohl, 2007, pp. 4–5).

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

#### 3.1 Stationarity

Before estimating a VAR model, the time series data must be checked for stationary. Accordingly, the means and variances are constant over time and the dataset does 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.

First, we check for stationarity for the twitter data (variable name: `nr_of_tweets`). Because this variable may not be stationary we also check for the first difference (`df_nr_of_tweets`) and the logarithmic first difference of this variable (`log_df_nr_of_tweets`). The results of the augmented Dickey-Fuller test are listed in the following table:

<code>dfuller nr_of_tweets, lags(0)</code>				
<code>dfuller df_nr_of_tweets, lags(0)</code>				
<code>dfuller log_df_nr_of_tweets, lags(0)</code>				
Dickey-Fuller test for unit root				
———— Interpolated Dickey-Fuller ————				
	Test Statistic	1% Critical Value	5% Critical Value	10% Critical Value
<code>nr_of_tweets</code>	-3.152	-3.475	-2.883	-2.573
<code>df_nr_of_tweets</code>	-14.167	-3.475	-2.883	-2.573
<code>log_df_nr_of_tweets</code>	-17.583	-3.475	-2.883	-2.573

Figure 3: Twitter variables stationarity test

From these result we can see that only the first variable is not stationary. This is because the absolute value of the Test Statistic (3.152) is smaller than the absolute value of the 1% critical value (3.475). The second variable is stationary ( $14.167 > 3.475$ ) and from here onwards we

are going to continuous the analysis by using this first difference transformation, which fulfils stationarity.

Secondly, we have to check for stationarity for the second variable, the bitcoin prices (variable name: bpi\_closing\_price). Here we also included the first difference (variable name: df\_bpi\_closing\_price) and the logarithmic first difference (variable name: log\_df\_bpi\_closing\_price) transformations. The ADF test for these variables gives the following results:

```

dfuller bpi_closing_price, lags(0)
dfuller df_bpi_closing_price, lags(0)
dfuller log_df_bpi_closing_price, lags(0)

```

Dickey-Fuller test for unit root

	Test Statistic	1% Critical Value	5% Critical Value	10% Critical Value
bpi	-2.609	-3.475	-2.883	-2.573
df_bpi	-16.303	-3.475	-2.883	-2.573
log_df_bpi	-15.996	-3.475	-2.883	-2.573

Figure 4: Bitcoin variables stationarity test

These result also suggest to use the first difference transformation of the variable bitcoin prices, for stationarity to be fulfilled.

## 3.2 Lag-Specification

To select the optimal number of lags to use for the VAR regression, we have to check for various information criteria. Information criteria are measuring the tradeoff between model fit and parsimony, giving the optimal number of lags to use (Brandt and Williams, 2007, p. 27). The calculations of these criteria for our specific data set are shown in the following table.

varsoc df_nr_of_tweets df_bpi_closing_price, maxlag(7)								
Selection-order criteria								
Sample: 12-20-2017 04:00:00 - 12-28-2017 10:00:00								
Number of obs						=	199	
lag	LL	LR	df	p	FPE	AIC	HQIC	SBIC
0	-3266.9				6.4e+11	32.8533	32.8667	32.8864
1	-3252.58	28.654	4	0.000	5.7e+11*	32.7495*	32.7897*	32.8488*
2	-3249.55	6.0623	4	0.195	5.8e+11	32.7593	32.8262	32.9247
3	-3247.69	3.7128	4	0.446	5.9e+11	32.7808	32.8746	33.0125
4	-3244.1	7.1874	4	0.126	5.9e+11	32.7849	32.9054	33.0828
5	-3238.16	11.868*	4	0.018	5.8e+11	32.7654	32.9128	33.1295
6	-3237.12	2.0852	4	0.720	6.0e+11	32.7952	32.9693	33.2254
7	-3235.03	4.1679	4	0.384	6.1e+11	32.8144	33.0154	33.3109
Endogenous: df_nr_of_tweets df_bpi_closing_price								
Exogenous: _cons								

Figure 5: Lag Specification

As can be seen, STATA calculates and presents various information criteria (AIC, HQIC, SBIC). The asterisk indicates the optimal lag length to use for the regression analysis. In this case all criteria suggest a lag length of 1.

### 3.3 Regression Model

After determining the optimal lag length the next step is to build the vector autoregression model (VAR model).

For this analysis, we use a basic unrestricted VAR model which consists of two endogenous variables,  $T_t$  which represents the variable *df\_nr\_of\_tweets* in our data set and  $B_t$  which represents *df\_bpi\_closing\_price*. As defined in the previous section the selected optimal time lag is 1. Given these information, the model equations can be written as follows:

$$\begin{aligned}
 T_t &= c_1 + a_{11}T_{t-1} + a_{12}B_{t-1} + \epsilon_1 \\
 B_t &= c_2 + a_{21}T_{t-1} + a_{22}B_{t-1} + \epsilon_2
 \end{aligned}
 \tag{1}$$

This model shows that the current value of our endogenous variables is determined by its own past values, the past values of the other variable and an error term. In our research approach we ask the question, if the amount of tweets mentioning bitcoin in the past, has an effect on the bitcoin prices in the periods that follow. This question is reflected in the second equation of this model.

After running the above VAR regression in STATA we get following results:

VARIABLES	(T <sub>t</sub> )	(B <sub>t</sub> )
	df_nr_of_tweets	df_bpi_closing_price
L.df_nr_of_tweets (T <sub>t-1</sub> )	0.00999 (a <sub>11</sub> ) (0.0693)	0.00463 (a <sub>21</sub> ) (0.0104)
L.df_bpi_closing_price (B <sub>t-1</sub> )	-0.675 (a <sub>12</sub> ) (0.469)	-0.125* (a <sub>22</sub> ) (0.0707)
Constant	-88.34 (c <sub>1</sub> ) (177.9)	-18.48 (c <sub>2</sub> ) (26.79)
Observations	205	205

Standard errors in parentheses  
\*\*\* p<0.01, \*\* p<0.05, \* p<0.1

Table 1: Var Regression

The equations of our VAR model are displayed vertically in this table. As said before the relevant equation in the model is the second one which is listed on the right column (B<sub>t</sub>). The coefficient of 0.00463 (a<sub>21</sub>) implies that a change of tweets by 1 tweet, changes the bitcoin price by 0.005 \$. This effect is very small and statistically insignificant (p value > 0.1). This answers our research question with the following statement:

The amount of tweets mentioning bitcoin has no significant effect on the price development of bitcoin.

### 3.4 Granger Causality

To assess the reliability (causality) of the previous results we use the Granger causality test. This test confirms if the statements by an estimated regression are valid. The following table shows the results of the Granger causality test for our VAR regression:

**vargranger**  
Granger causality Wald tests

Equation	Excluded	chi2	df	Prob > chi2
df_nr_of_tweets    df_bpi_closing_~e		2.0732	1	0.150
df_nr_of_tweets	ALL	2.0732	1	0.150
df_bpi_closing_~e    df_nr_of_tweets		.19677	1	0.657
df_bpi_closing_~e	ALL	.19677	1	0.657

Figure 6: Granger causality test

The relevant results for our regression are displayed in the lower row. The Null hypothesis here is that the lagged (lag 1) values of the number of tweets does not cause a change in the bitcoin prices. The Alternative hypothesis would then be that the lagged values of the number

of tweets does cause a change in bitcoin prices. As the results show, the Probability value is 65.7% which is clearly more than 5% and constrain us to accept the Null hypothesis and conclude that the regression results stated above are valid.

## 4 Conclusion

Understanding, describing and predicting the current phenomenon of cryptocurrencies has been the objective of many researchers. A variety of methods and approaches have been applied and mixed results have been produced. Our goal for this seminar paper was to investigate the possibility of the existence of a correlation between twitter activity regarding bitcoin and the price evolution of bitcoin. After collecting hourly data for approximately 8 days we generated 2 time-series datasets and decided to perform a statistical analysis by using the Vector Autoregression (VAR) approach.

According to our regression results, changes in twitter activity due to more or less tweets about bitcoin have a very small and insignificant effect on future changes of bitcoin price. The validity of this statement was also verified and confirmed by the Granger causality test. At this point, however, it must be pointed out that only a small period of time has been considered in this analysis and the results cannot be generalized for the overall trend of cryptocurrencies in the recent time or for the future. We are also aware that we may have observed a period of abnormal behavior of bitcoin prices compared to the trend of the last months.

We encourage further research in the same direction as this paper and would be interested to see an analysis with observations over a longer period.

## 5 References

- Brandt, P. T. & Williams, J. T. (2007). *Multiple time series models* -. London: SAGE.
- 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.
- Lütkepohl, H. (2007). *New introduction to multiple time series analysis*. Springer Berlin Heidelberg.
- 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 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.

Alen Stepic - 11-475-258

Dimitrios Koumnakes - 10-613-370

Severin Kranz - 13-606-355

Joël Sonderegger - 11-495-488

Chi Xu - 16-300-915

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.