08/12/2017

Network Dataset

MSING055: Programming for Business Analytics

Group Project: Vensel Csergö (17104337), Taline Filipovic (14003685), Soumaya Mauthoor (20069045), Jakub Kneppo (14960908), Luana Totea (14050715)

Dataset for Analysing Patterns of Insider Trading

Word Count: 2088

Table of Contents

1. Executive Summary 1
2. Introduction 1
3. Methodology 2
4. Analysis 4
   - 4.1. Time Series Analysis 4
   - 4.2. Network of insiders 6
5. Identified limitations with data sourcing 7
6. Future Recommendations 7
   - 6.1. Network Analysis of Insiders 7
   - 6.2. Opportunistic Insiders 8
7. Conclusion 8
8. References 9
9. Appendices 10

Appendix A 10
Python Script 10
Appendix B 10
- SEC Form 4, Insider Buying & Selling Data 10
- Google Finance Historical Stock Price Data 10

Appendix C 11
- Summary of the Dataset Statistics 11
- Buy and Sell Deciles 11

Appendix D 12
Members' contribution, overall performance and project reflection 12
- Work breakdown structure and contribution 12
- Assessment of the Overall Performance 13
- Reflection on the Project 13

1. Executive Summary

The project was structured into five incremental parts. Firstly, it introduced the context behind illegal insider trading and the opportunities associated with data mining insiders' filings and historic stock prices. Consequently, a new dataset was created using Python programming language in order to scrape, cleanse and merge U.S. insider transactions with stock price data. Additionally, a binomial time-series analysis was conducted on the newly created dataset where graphic visualisations were generated using Tableau software. The paper argued that understanding the temporal patterns of insiders' trading behaviors, anomalous activities serve as support for regulators to help combat financial crime. Subsequently, this study built on the concept of networks by identifying a network of insiders derived from the new dataset to help elicit the dynamics of trading behaviour between individuals. Finally, it concluded with the summary of the work and recommendations to further leverage the usage and reproduction of the new dataset for further investigations.

2. Introduction

The U.S. Securities and Exchange Commission (SEC) defines legal insider trading as highly ranked corporate insiders who trade securities issued by their own corporation (SEC, 2017). It is recognised that a quarter of all public deals involve some extent of insider knowledge[1] which leads to the notion of *illegal* insider trading (Dealbook, 2017). Thus, the SEC introduced Form 4 documentation that records all buy and sell transactions of corporate insiders, which is publicly accessible, allowing for better detection and understanding of trading activities. Jointly, they cooperate with investigators from the Financial Industry Regulatory Authority (FINRA), the Options Regulatory Surveillance Authority (ORSA) and NYSE Regulation to investigate suspicious trades and tackle insider trading.

Within this context, *Advances in Social Network Analysis and Mining* (ASONAM) published a paper on the discovery and analysis of insider trading connections that would aid legal bodies to identify viable litigation cases for unlawful trades. Drawing upon Tamersoy et al's (2013) work, this paper will seek to identify patterns and factors amongst traders that incentivise buy and sell transactions, connections between insiders and the profitability of trades by merging Form 4 filings with respective stock price changes.

[1] Insider knowledge - allegedly obtained non-public information about a particular stock investment. This information is only available to high level executives or shareholders in the company known as corporate insiders.

3. Methodology

A dropbox link to the Python code for all modules used to construct the dataset is provided in Appendix A. To begin with, the main dataset was scraped from Form 4 SEC filings with all information published at the url provided by Appendix B: SEC Form 4, Insider Buying & Selling Data. As there was no API available, HTML scraping was used instead with the help of the beautifulsoup library (file: "1_main_data.ipynb"). Admittedly, the website was sorted according to the date of acceptance of SEC Form 4 as opposed to the actual transaction date. Thus, in order to obtain all trades transacted in between 2014 and 2016, the data accepted between 2013.01.01 to 2017.11.02 was scraped. Subsequently, a separate filtering algorithm was used in order to select just the insider trades transacted in between 2014 and 2016 only (file: "2_filtering.ipynb"). This resulted in the creation of the primary file with 299,572 rows of transactions and associated 11 variables used for the merged dataset.

The second step was to cleanse the data scraped from SEC Form 4 filings, in particular the transaction stock index symbol or "ticker" (file: "3_cleansing.ipynb"). Many of the transactions had tickers in different formats, which made it difficult to compare them to the actual stock market tickers. Thus, the data was filtered out for validity checks and all spaces, brackets and other special characters were removed from the tickers.

The third step was to collect the secondary dataset to be merged which consisted of the open stock price of all traded shares 1 week post the transaction date and time. This information was used to determine how profitable each trade was based on the one week change in every stock price. The data was scraped from Google Finance (Appendix B: Google Finance Historical Stock Price Data) based on the trade symbol and the transaction date (file: "5_secondary_data.ipynb"). Admittedly, the initial solution to mine the corresponding share price, row by row for all 299,572 transactions was very slow with an estimated execution time of 58 hours. In order to solve this issue, a grouping method was used instead so to optimise the runtime (file: "4_grouping.ipynb"). In order to implement this method, all the unique combinations of trade symbols and the date time, seven days post the transaction (i.e. the "Week_Later" variable), were collected. Then, for each of the symbols, all the "Week_After" dates were transposed into a list of variable length. This translated into all stock data corresponding to 1 trade symbol to be downloaded once only, retaining only the data that matched the list of possible dates. The alternative method resulted in accessing Google Finance only 5,800 times to get the estimately 5,800 trade symbols in the original filtered and cleansed dataset, compared to the initial method, where the written logic was parsing the website 300,000 times. In short, this alternative solution resulted in cutting down the scraping time from the initial 58 hours to just 35 minutes.

Lastly, the secondary dataset, containing the open stock price a week after the transaction date was merged with the primary dataset ("6_merging.ipynb") by using a left join dataframe object based on the transaction symbol and "Week_After" date. This resulted in the creation of the final dataset for the analysis discussed in Section 5.

It is worth mentioning in the methodology that the working group took a proactive approach in conducting a safety analysis of the integrity of the dataset to be used for analysis in Section 5. As aforementioned, Google Finance web service was used to obtain the historic price data due to the relative ease of use. Admittedly, the working group recognised the associated data integrity risks of Google changing this particular API in the past and that it did not officially support it. Therefore, in order to mitigate this risk, historical price data from well acclaimed web services such as Yahoo Finance or alphaadvantage.com was extracted and compared against the group's findings. It was propitious to learn that all sets of data were identical in figure, thus no time constraints hindered the working group's data analysis.

The approach the team took to divide the workload for the methodology outlined in four intrinsic parts, was to to divide the tasks into six different modules where the data was transferred from one module to another via csv files. The modules were stored in 'SherlockML', a cloud based environment for ease of storing, manipulation, testing and analysis which enabled the team to work on different modules simultaneously and in seamless collaboration. A summary statistics of the data is provided in Appendix C: Summary of the Dataset Statistics

4. Analysis

4.1. Time Series Analysis

The first analysis undertaken is a time series across three years, which visualises 'who trades when'. First approach seeks to question whether trades are regular, i.e. influenced by the outside environment, whereas the second approach depicts the influence of insider position on transaction activity.

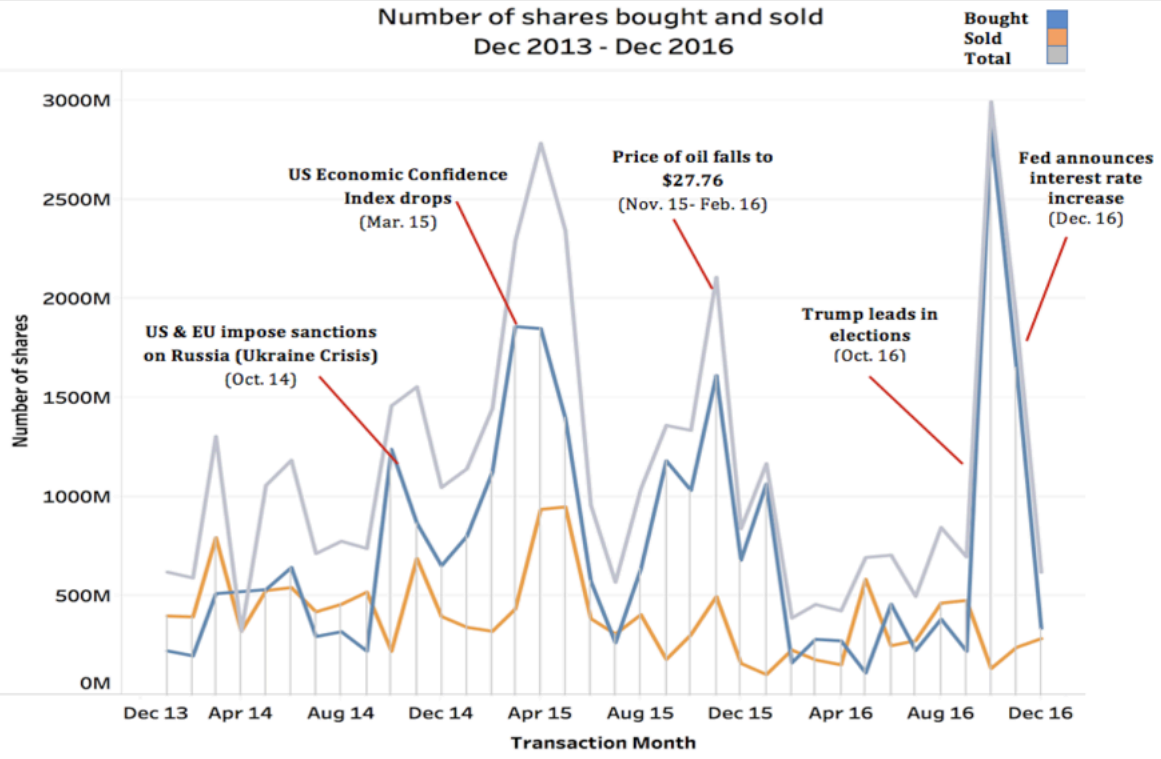

Figure 1: Daily transactions of buying and selling stocks conducted by insiders through December 2013 and December 2016. The pattern indicates more 'buys' prior to elections where confidence in candidates of interest rises (2016 election). Inversely, 'selling' positions increase during detrimental global political and economic events (Ukraine crisis and Oil price fall).

Based on Figure 1, it can be concluded that the dataset illustrates a rather standardized approach overall amongst insiders and the timings of undertaken trades. Hence, there is a role in outside climate that affects behaviour and potential of profit-making. Simply, traders either profit by buying cheaper prior to events relating to economic growth or minimise loss by liquidating before 'crashes' at current higher prices. Moreover, using this analysis, we further explore those insiders (in terms of positions) that are highly affected by the outside environment. Those who have larger trade volumes often carry more information (either public, private or both) in order to profit more by the events.

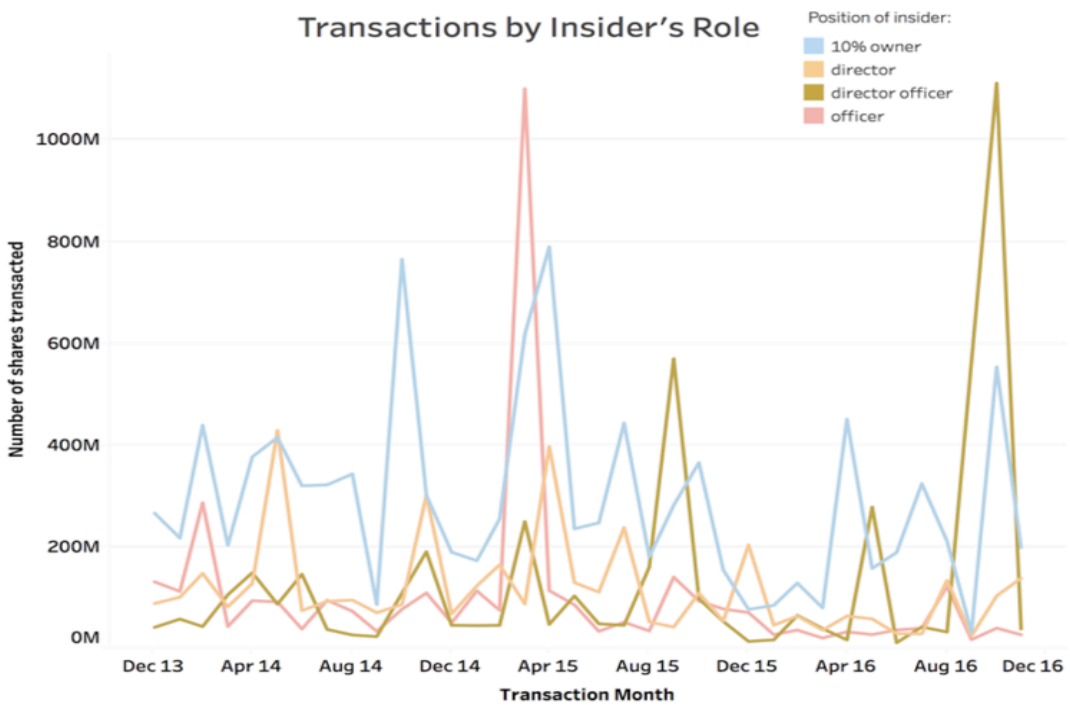

Figure 2: The number of shares transacted by the role and positions of insiders. The trading differences over time analyses and distinguishes between large shareholders (owning approximately 10% of firm's shares), directors, officers and those who hold multiple positions (active in more firms), such as director-officer.

The largest shareholders (10% owners) show higher numbers of activity and transaction volumes throughout time (due to amount possessed), whereas director-officer tend to have sharp spikes in transactions when events are 'domestic' (US elections and US fall in domestic oil prices). Overall, there is a clear activity pattern in transaction amounts by insider that is consistent with economic and political events from Figure 2, which influence the profitability of shares and transaction activity. The closest findings to irregular behaviour would be the spikes in April 2015 from 'officer' and December 2016 from 'director-officer'. To gain more understanding into anomalies, our dataset can be used to further dissect firms and respective individuals were involved in transactions at that specific time. It was found that they were large liquidations of individuals leaving their firms. Hence, this is not flagged as an unusual activity.

4.2. Network of insiders

The working group aimed to build a network amongst insiders for a firm arbitrary selected from the merged dataset with a view to identify possible communication tunnels.

For the purpose of their micro study, the working group created a network of all insider trades of '1347 Capital' firm as illustrated in Figure 3. Expectedly, the network's structure resembled a 'small-world network' (Newman and Watts, 1999) in which the nodes (insiders) were not all mutually connected, but shared mutual neighbours.
The working group aimed to investigate whether the connections (edges) denoted a similarity in trading patterns between the nodes within 1347 Capital. Provided that the resulted network displayed many mutually shared 'neighbours', that could have indicated a presence of mutual patterns of insiders who continuously traded on same dates for further behavioural analysis.

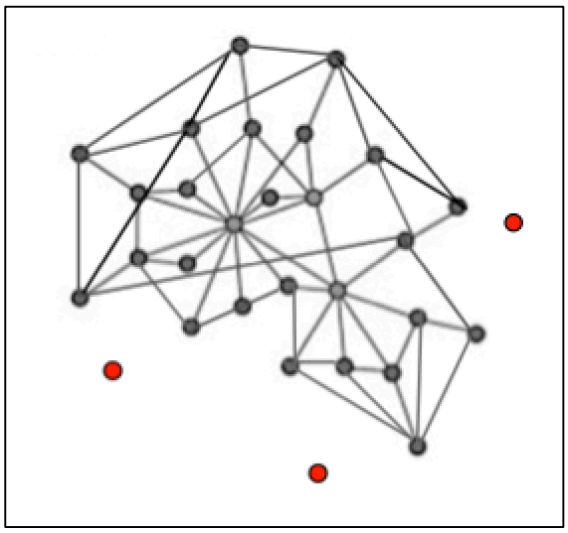

Figure 3: Network of 29 individuals and presence of mutual trading behaviour of all securities of '1347 Capital' firm constructed using Gephi software.

The illustrated network showed that three of positioned individual had no sign of mutual trading behaviour with others, however an interesting view was that there were two nodes (traders) that acted as 'hubs'. Essentially, the working group noted that there were many individuals held who mutual connections, and if they were taken out from the network, the connectivity would have been substantially lower. This meant that two traders had higher trading activity consistent with other individuals. However, the working group acknowledges that these observations do not necessarily imply that higher ranked individuals trading behaviour was 'imitated' by those beneath. Further network centric analysis needed to be conducted to determine whether those two insiders were similar enough in terms of trading behaviour.

5. Identified limitations with data sourcing

When scoping the project, limitations were encountered with data access approvals, the presentation of values as well the overall consistency of the merged structure of the dataset. To begin with, it was acknowledged that there were differing formats relating to several fields, i.e. the price and number of shares traded were at times either empty or had invalid value formatting (millions Vs. billions $). Because of the fact that the original files were missing critical data, the percentage change in stocks' price was allocated to only 75% of the trades. Nevertheless, despite rigorous cleaning of the data, the working group recognised that persistent faulty fields could have potentially affected and skewed the statistical distribution, and ultimately the results of the analysis.

Additionally, 10% of the Buy insider transactions showed abnormally large increases in stock price that could have potentially skewed findings in Section 4. (Appendix C: Buy and Sell Deciles). The working group attributed this data sourcing limitation to insiders' opportunity to buy trades at nominal values in lieu of bonuses.

Lastly, the initial intent of the study was to include the U.K. financial market to further differentiate the approach used in this paper from Tamersoy et al's work (2013). However, this would have required consent from the Financial Conduct Authority (FCA) and data mining was therefore limited to published insider data for the U.S. financial market.

6. Future Recommendations

6.1. Network Analysis of Insiders

In Section 4, the working group discussed their initiative to build a network amongst insiders of the transactions stored in the dataset for 1347 Capital. However, in order to determine similarity in insiders' trading behaviour, they recognised the need for a rigorous network centric analysis to be conducted in the future .

One possible approach would be to apply a similarity function (Formula 1) such as the one defined by Tamersoy et al (2013, p. 802) in conjunction with using the percentage change in stock prices to determine the profitability of trades.

$$S(X_C, Y_C) = \frac{\left(\sum_{i=1}^{|X_C|} \sum_{j=1}^{|Y_C|} I(x_i, y_j) \right)^2}{|X_C| \times |Y_C|}$$

Formula 1: Similarity function of transaction dates where X_c and Y_c denote the set of transactions of company C by insider X and insider Y. Additionally, xi and yi in the numerator denote the sum of traders' X and Y transactions.

If the formula would derive a result of 1, there would be a strong connection/pattern between insiders of 1347 Capital, whereas 0 would indicate absence of similarity. The formula could analogously be applied to all insiders in the firm, which would give a full cross-comparison of all employees at the firm.

It is worth mentioning that even though the network of insiders at 1347 Capital was dense and applying the aforementioned formula could derive many mutual trading patterns, an additional step would be required for a rigorous analysis. Because calculating density does not necessarily indicate illegal communication of insider information, percentage stock price changes from the merged dataset could be analysed to furtherly filter connections between insiders. Essentially, this would flag the connection between those individuals who show consistent above average profits derived from their trading actions. If noticeable, this could reinforce the communication of privileged information in the network and ultimately will enable the ability of detecting opportunistic insiders, and flagging them.

6.2. Opportunistic Insiders

So far, the working group considered creating a network of insiders' whose trades coincide on same dates within 1347 Capital firms. Another further dimension that can be facilitated by the unique dataset is to create a network of insiders' with yearly profits showcasing abnormal returns (that are consistent with price change) in order to capture a file of opportunistic insiders to be further investigated. It would be worth reproducing these results and analysing whether the same trading behaviour is consistent throughout firms in specific sectors or if it is unique to a firm.

7. Conclusion

In conclusion, this paper gives a novel insight into insider trading patterns. It draws on Tamersoy et al's (2013) work by offering a a juxtaposition of time series analysis and network analysis. However, and most importantly, it enhances on the analysis by adding a new dimension focused on the profitability derived from insider transactions at a given time to support investigators document anomalous activities on a large scale. The goals and objectives set by the working group for this project were not to find illegal traders, but rather document patterns, irregular behaviour and other trends that make sense of the dataset and its dynamics.

8. References

Nasdaq.com. (2017). SEC Insider Form 4 - NASDAQ.com. [online] Available at: http://www.nasdaq.com/quotes/sec-insider-form-4.aspx [Accessed 5 Dec. 2017].

Newman, M. and Watts, D. (1999). Renormalization group analysis of the small-world network model. Physics Letters A, 263(4-6), pp.341-346.

Sec.gov. (2017). SEC.gov | Insider Trading. [online] Available at: https://www.sec.gov/fast-answers/answersinsiderhtm.html [Accessed 7 Dec. 2017].

Sorkin, A. (2017). Study Asserts Startling Numbers of Insider Trading Rogues. [online] DealBook. Available at: https://dealbook.nytimes.com/2014/06/16/study-asserts-startling-numbers-of-insider-trading-rogues/?r [Accessed 7 Dec. 2017].

U.S. Securities Insider Trading Information. (2017). Insider Trading - Insider Trades, SEC Form 4, Buying & Selling Data. [online] Available at: http://insidertrading.org. [Accessed 7 Dec. 2017].

Wutkowski, K. and Younglai, R. (2017). Regulators band together to fight insider trading. [online] Available at: https://www.reuters.com/article/us-insidertrading/regulators-band-together-to-fight-insider-trading-idUSN1525923720070815?dlbk2 [Accessed 7 Dec. 2017].

9. Appendices

Appendix A

Python Script

The python code for all modules discussed in the Methodology can be accessed through the following dropbox link-
https://www.dropbox.com/sh/e25gjwbh0fj215i/AABgepoTAThsKEtvv3WPF6bya?dl=0

Appendix B

SEC Form 4, Insider Buying & Selling Data

The Form 4 dataset contains information on all insider trade transactions between 1986 and 2017 such as: whether it was a buy or sell, date of transaction, date of acceptance, issuer name, trading symbol of the stock, reporting owner, reporting owner relationship, number of shares, price/share, total value of trade, shares owned following the transaction. (U.S. Securities Insider Trading Information, 2017) The form can be found on the following URL: http://insidertrading.org

Google Finance Historical Stock Price Data

Google finance provides historic stock price data. The desired trade ticker and dates are specified in the url:
https://finance.google.com/finance/historical?q=SCTY&startdate=%272014-06-17&enddate=%272014-06-24

The data can be outputted to csv by appending $output=csv to the url.

Appendix C

Summary of the Dataset Statistics

Number of transactions	299,572
‘Buy’ transactions	122,878
‘Sell’ transactions	169,694
Number of unique insiders	48,061
Number of unique companies	6,252
Number of unique trade symbols	5,889
Trades with calculated stock price % change	225,408

Buy and Sell Deciles

Decile	From	To
10%	-99.9	-6.403
20%	-6.403	-1.714
30%	-1.714	4, 0.0
40%	0	438
50%	1.438	3.964
60%	3.964	15.145
70%	15.145	62.483
80%	62.483	163.426
90%	163.426	542.207
100%	542.207	69,949,900

Decile	From	To
10%	-100.001	-7.488
20%	-7.488	-3.428
30%	-3.428	-1.788
40%	-1.788	-0.682
50%	-0.682	0.224
60%	0.224	1.176
70%	1.176	2.364
80%	2.364	4.348
90%	4.348	14.736
100%	14.736	999,999,900

Appendix D

Members' contribution, overall performance and project reflection

Work breakdown structure and contribution

After preliminary discussion on what how we want to go about our paper, the group was divided in order to leverage strengths of the individuals. Three members worked on the coding, scraping, cleaning and merging of the data, while two members focused on the analytical analysis of the paper, formulating approaches to achieve results, and prepare the write up of the paper. The specific breakdown is further explained in the table below:

Soumaya Mauthoor	Investigated multiple approaches and wrote python scripts to obtain the stock index price one week after the transaction date: 1. Using the googlefinance.client python library https://pypi.python.org/pypi/googlefinance.client 2. Scraping the html in google finance webpage using beautifulsoup 3. Looping through a csv output from the google finance webpage Wrote python script to merge secondary data set with primary dataset and obtained some summary statistics. Assisted with the formulation of the network analysis
Taline Filipovic	Took part in forming the initial idea of the project, its approach and data sources. Highest involvement was within the analysis part, which other than writing it, involved constructing the visualisations on Tableau and trying to tackle network analysis (unfortunately I was unable to realise my initial idea of a larger network). Essentially, I really enjoyed working with the group, the dynamic was great and everyone truly contributed to their full ability.
Luana Totea	I was involved in the front to back write up of the project, making sure our scope was clearly defined from one corner to the other. Additionally, I supported Taline in the formulating the creation of the Network and in brainstorming ideas of how to best leverage the merge dataset.
Vencel Csergő	Wrote the code to scrape the main dataset used in the study. Also assisted with scraping the secondary dataset and cleaning it and wrote the methodology part of the final write-up.
Jakub Kneppo	I was involved in programming part of the script (I was responsible for 'cleansing' the data), contributed to writing the methodology and also took part in forming the initial strategy. Overall, working in this group was a very positive experience and everything was great in terms of the group dynamic. Personally, I felt like a valued member of the team and I learnt a lot about data analytics and significantly improved my Python skills. Overall, I think this was a positive experience and I would say that there was nothing overly negative.

Assessment of the Overall Performance

The performance of the group was highly recognised amongst all involved individuals. Every single person was involved from the start of the process, which began already two days after group allocation. There was strong understanding between individuals' stronger skills in respect to specific parts of the project, but regardless of the division of skills, it was important for the team to have everyone interact from start to finish. There was strong communication throughout, so even members who did not scrape the data or write the network implications, can nevertheless explain it. The ongoing communication was crucial to meeting team milestones and delivering the final draft on time.

Reflection on the Project

When reflecting on how the project went, there was an overall positive experience in terms of meeting our final expectations. The project was started early and this was an advantage that we leveraged in order to 'fix' anything that did go wrong. More notably, our merging of the data and computational aspect of stock price changes was a rather difficult task to tackle, so time management was prioritised in order to effectively execute this aspect. However, there was difficulty in constructing larger networks of different firms that are connected by mutual individuals, which was rather too complex to realise. As well, the selection of firm for the network analysis was based on the notoriety of the firm within the domain of trading, as it was difficult to select a company that would provide reliable insight into the communication of non-public information, illegal trading. Overall, there is still available scope and depth of analysis to be undertaken that was constrained by possessed expertise in analysing networks to greater extents. Ultimately, what the project provides is a solid idea on how our new dataset can be used for the purpose of analytical insight.