Team#037 - Pairs Identification for Pair Trading: A Data Viz Approach

{smeduri7, pbaikani3, sv67, reshed6, nbadri3, pnarula6}@gatech.edu

1 INTRODUCTION: STATISTICAL ARBITRAGE - PAIR TRADING

The strategy behind pair trading is to identify a pair of securities (or baskets of assets) whose price tends to move together, and once it is identified the trader tracks the spread (the difference between the assets prices) and once the spread widens, long (buy) the asset whose price decreased and short (sell) whose price increased, assuming the assets would maintain their relationship and the prices will revert to their historical trend. While there are many methodologies to test pairs for co-movement, the two most common are the minimum-distance method [2, 6, 8, 11] and the cointegration method [4, 6]. Pair trading requires three steps - Pairs formation, co-movement relationship testing and determination of trading entry and exit rules [7]. This project is mainly focused on the first two steps.

2 PROBLEM DEFINITION

An initial pairs formation must be performed before actual comovement is tested. The purpose of the project is to serve as a proof of concept of finding all worldwide potential pairs/baskets of assets using the "cheapest" computational method while eliminating any subject matter expertise required to perform such initial pairs formation. Starting with a list of all U.S. stocks, we find which companies are mentioned in financial news together with a company selected by the user and present the results using a network graph [3, 16]. The assumption is that companies mentioned in the same news articles are potential pairs for comovement testing.

3 SURVEY

Currently, the pairs formation process begins with running comovement tests on all possible pairs and eliminating pairs that fail the tests [2, 7]. This method is inefficient, computationally expensive and not feasible for testing stocks within different markets. Another pairs formation method is rule-based which is subjective and requires industry and market expertise [17, 15]. We plan to develop a cross-market pairs formation model that is efficient and eliminates industry, domain and market expertise required by the currently used methods. We believe the initial pairs formation is a crucial step in pairs trading and we are optimizing it to identify potential pairs. We also enable comovement testing steps to be more efficient by significantly reducing the search space of the pairs to test. We'll be using the finance media articles data [1] to find the co-occurrences of the companies and build the companies co-occurrence network [3, 5, 16, 18]. It's a proven method for the social network generation using the Reuters news articles [10] which demonstrated various techniques to calculate the connection weightages, communities (stock baskets/buckets in trading domain). We believe that applying the same to pairs trading problem would be successful. Additionally, visualize pairs of highly correlated stock listings and related metrics like stock data comparisons, co-integration metrics [9, 12] etc.

4 PROPOSED METHOD

4.1 Intuition - why should it be better than the state of the art?

Finding cointegrated stocks to trade on either requires market and industry knowledge or entails an inefficient and expensive process of testing all possible combinations of stocks for cointegration. Our model, focused on initial pair-formation, will eliminate the need of any knowledge required to form pairs for testing and make the actual testing step more efficient by significantly reducing the number of pairs to test.

4.2 Description of approaches

4.2.1 *Data Collection*

The news article data is collected using an API offered by *newsfilter.io*. The key feature we are interested in is the co-occurrence of company names in news articles. The news data for the past 2 years is collected for each company in an AWS instance and stored in a S3 bucket. Later this data is processed and stored in a SQL database. The other data we require is the historical market prices for each company, this data is being fetched at run time using the *yfinance* python package. Additionally, a complete list of U.S. tickers using a free API was retrieved from EOD Historical Data.

4.2.2 *Data Storage*

There are around 19k stock assets (identified by tickers) being visualized in our graph network. For each stock asset, financial news articles that mention the name of the stock asset were collected for the past 2 years. There are around 5 million news articles collected and stored on AWS S3 storage. Each article is processed and stored in a structured relational database with the following schema:
"`article_id, ticker#1, ticker#2, publication_date`"
where `article_id` is the id of a finance news article. `ticker#1` is the stock asset associated with the news article and `ticker#2` is the stock asset co occurring with `ticker#1` in that news article published on `publication_date`. The adjacent nodes for a node (stock asset `ticker#1`) in the graph network are the stock assets highly co-occurring with that node's stock asset `ticker#1`. It can be found by filtering all the articles for `ticker#1` and grouping them by `ticker#2` and sorting by the number of `ticker#2` records. The top 6 tickers are selected and edges are formed in the graph network. Retrieving news data for the past 2 years, the dataset contains $\approx$ **29 million** (2.6 GB disk size) `ticker#1-ticker#2` co-occurrences.

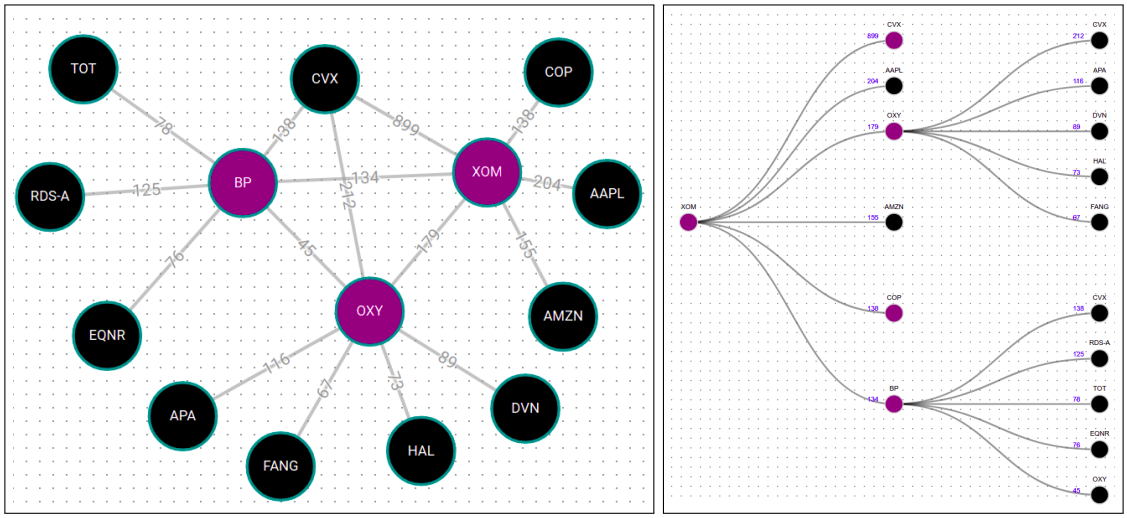

Figure 1—*(Left) Ticker Network Graph; (Right) Ticker Network Tree*

4.2.3 *Ticker Network Graph*

The network graph of the company tickers is constructed from the data obtained after processing the raw news filter API results [1]. The nodes of the network are the company tickers and an edge is formed between two nodes if they appear in the same article. The edge weight corresponds to the number of times the two tickers were co-mentioned in the articles [10]. By starting with an initial ticker, we expand the network by a branching factor of b (≈ 6) constructing b edges with highest weights by considering the co-occurrences of the tickers in the news articles until a depth based on the user interactions with the graph [3, 5, 16, 18]. The user can further select from the displayed tickers any two

nodes to view line charts comparing stock prices, spread line, cointegration "score" and correlation which will be presented analyzing historical stocks prices within a timeframe determined by the user. Another visualization the user will be able to utilize to understand companies' co-occurrence in the news articles is an expandable hierarchical tree, with the number of co-occurrence appearing next to each edge connecting two nodes.

4.2.4 *Correlation/Cointegration Test*

1. Pairwise Cointegration Test: To test pairs for cointegration we used the Engle-Granger test which is a regression based approach that states that given two ***non-stationary*** series X_t and Y_t, if there exists α and β such that the residuals (which are referred to as spread) of the linear combination $Y_t = \alpha + \beta X_t + \epsilon$ are stationary, then the variables X_t and Y_t are said to be cointegrated and the state of $Y_t - X_t - \alpha$ is stationary. Note that the linear combination above can be written as $\epsilon = Y_t - \beta X_t - \alpha$. The inclusion of the intercept should be carefully considered. Whether to include it or not depends on the subject matter tested for cointegration and the analyst's own consideration. Since the intercept term will not be traded upon and merely shifts the spread line upward or downward, we omit it [7, 13, 14].

By design, assuming the linear combination $Y_t = \beta X_t + \epsilon$ holds, the spread's mean is 0 and the spread line will revert to the mean if stationary. To test whether the spread is stationary, we perform the Augmented Dickey Fuller Test (ADF) with no constant and trend, and lag = 1 (note that Akaike information criterion (AIC) and Bayesian information criterion (BIC) can be used to get the optimal lag p, but for simplicity and due to the fact that including lags of order 2, 3, 4.... does not add much value, we keep the model with order 1)[7] which is formulated as $\Delta\epsilon_t = \delta\epsilon_{t-1} + \phi\Delta\epsilon_{t-1} + \nu_t$. The test is essentially a ***unit root*** test for stationarity where we would want to reject the null hypothesis that a unit root does exist ($H_0 \rightarrow \delta = 0$), and hence the spread is stationary and the pair of stocks are cointegrated. The intuition behind the test is that if the null hypothesis is NOT rejected (i.e. $\delta = 0$), the lagged level (ϵ_{t-1}) provides no relevant information about the current level change ($\Delta\epsilon_t$) besides the one obtained by the lagged change. However, if the null hypothesis is rejected, the lagged level does provide information about the current change and the series exhibits reversion to the mean of 0.

What is Stationarity? A time series has stationarity if a shift in time doesn't cause a change in the shape of the distribution. Basic properties of the distribution like the mean, variance and covariance are constant over time.

What is a Unit Root test? A unit root is a stochastic trend in a time series, sometimes called a "random walk with drift". If a time series has a unit root, it shows a systematic pattern that is unpredictable.

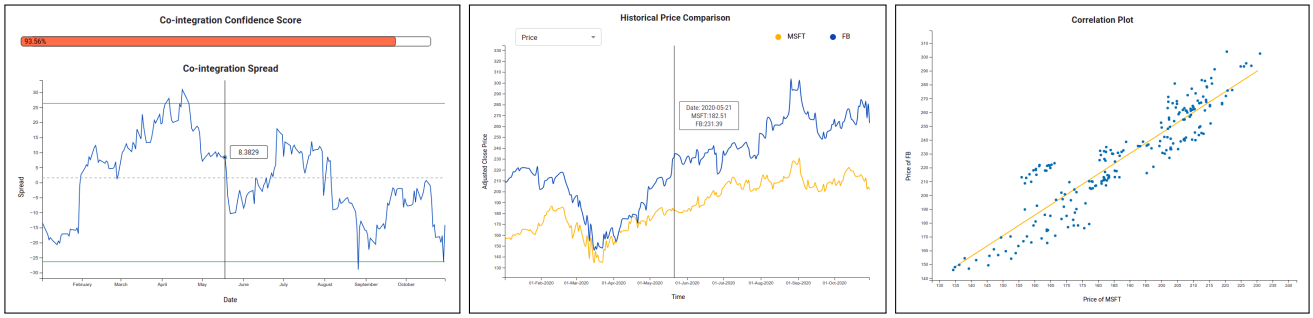

Figure 2—*(Left) Co-integration graphs; (Center) Price comparison chart; (Right) Correlation scatter plot*

2. Correlation: It is important to understand that cointegration is not correlation. Correlation is used to check for the linear relationship between two variables while cointegration is used to check

for the existence of a long-run relationship between two or more variables. In other words, using stock prices as an example, if two price series are positively correlated, it means that both series move in synchrony, as one stock goes up, so is the other stock. If the two stocks are cointegrated, the prices cannot "wander off" in the opposite direction for very long. They will have to revert back to an equilibrium state. Moreover, two time series can be both correlated and cointegrated, correlated but not cointegrated, or cointegrated but not correlated. The stock prices are regressed against each other to generate a scatter plot for the visualization of the linear relationship between the two stocks and to serve as a proxy for correlation.

4.3 List of Innovations

1. It is a novel approach in using financial news articles data to find the co-occurrences for pair trading. The current methods are based on financial/subject-matter expertise.
2. Ticker Network graph building from their co-occurrences in the financial news articles is an innovative thought by using the analogy of the social networking [10]. The intuition is that co-mentioned businesses/stocks are from the same industry, share similar risks, share some relationship such as buyer-supplier, etc.
3. The UI visualizations of the tickers, network graph, correlation and co-movement metrics are novel as it is a synergy of different kinds of datasets (news, stock data) used for pair trading.
4. The overall idea to make sense of the news data in a different domain of the stock market and take advantage of this information to eliminate subject matter expertise is an innovative thought.

5 EXPERIMENTS/ EVALUATION

Q1) How to collect a large volume of data and process at scale? We started the data collection using the newsfilter API locally, but the data collection script got interrupted multiple times due to network connectivity issues and requirement of high storage was also a challenge. So data collection was moved to an AWS instance and the data was stored in an S3 bucket.

Q2) What should be the network graph branching factor (b) and the expansion depth (d) to find the best combination resulting in a good correlation between the stock tickers? The depth was left to the user and the graph is progressively constructed based on user interactions. WIth the branching factor of 6, the relevant tickers are returned which are further validated with the co-movement metrics.

Q3) What factors to consider for Cointegration Tests? We evaluate results based on whether a strong cointegration exists within pairs of stocks mentioned in the same article. More specifically, we anticipate that the cointegration score of two companies is related to the number of co-occurrence. Meaning that the cointegration score of two stocks with co-occurrence value of 2,000 will be higher than the cointegration score of two stocks with co-occurrence value of 1,000. With that being said, it is important to note that the existence of strong cointegration (or lack of) should be carefully evaluated, because if cointegration exists when testing for the last 6 months, for example, it does not necessarily mean that it exists when testing for different time intervals (e.g. 12 months). Determining the time interval to check for cointegration is subjective and must be evaluated by the user based on his/her trading strategy. This is why the tool allows the user to select start and end date to test for cointegration.

Q4) Is the tool able to find "obvious" pairs for pairs-trading? As expected, the tool definitely indicates "obvious" pairs for potential cointegration. For example, when analyzing for Exxon Mobil Corporation (XOM), the tool indicates to potentially pair it with British Petroleum (BP). When testing for cointegration, we see that the two stocks are cointegrated with a score of 94.4% be-

tween the timeframe of 01/01/2020 - 11/01/2020 and 99.8% between 01/01/2019 - 11/01/2020.

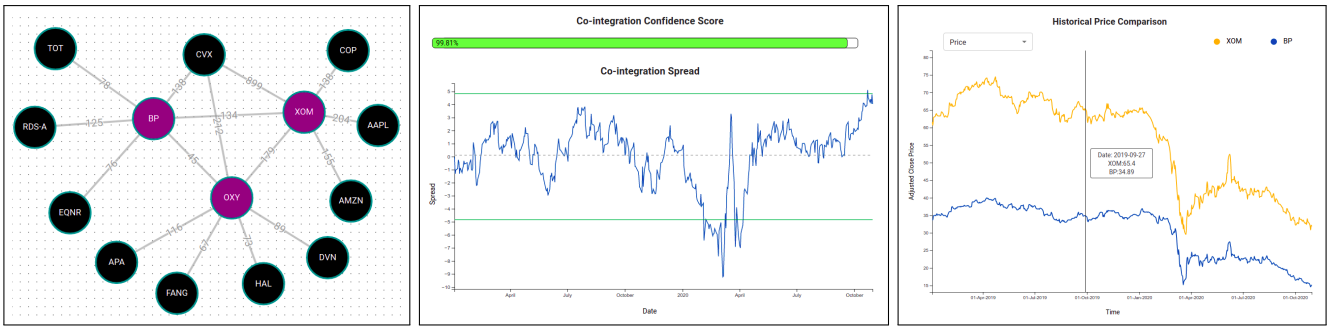

***Figure 3—XOM vs BP**: (Left) Network graph; (Center) Co-integration graph; (Right) Price plot*

Q5) Is the tool able to find cointegrated pairs that are most likely to be overlooked when testing using a rule-based approach? It makes sense to test for Apple and Amazon, but what about Apple and Tesla? To prove the tool achieves its goals, it indicates Apple (AAPL) and Tesla (TSLA) as a potential pair worthwhile to test for cointegration. While Apple's industry is Consumer Electronics, Tesla's is Auto Manufacturers and we believe traders might not consider the two as a pair to trade on. However, when checking for cointegration within 01/01/2020 and 11/01/2020, we get a strong cointegration (92.6% confidence level), suggesting the two stocks can be traded on using pair-trading strategy. Moreover, expanding the graph on Tesla, the tool indicates that General Motors (GM) should be tested for cointegration with Tesla. This suggests that it is also worthwhile to test for cointegration between Apple and General Motors, and indeed we find a strong cointegration score of 97%.

Another example is Gaming and Leisure Properties Inc (GLPI) and Atlantica Sustainable Infrastructure PLC (AY). GLPI is a gaming-focused real estate investment trust while AY is a sustainable infrastructure company that owns and manages renewable energy, efficient natural gas, transmission and transportation infrastructures and water assets. However, as indicated by the tool that the two stocks might be cointegrated, their cointegration score (within 01/01/2020 and 11/01/2020) is 98.4%.

6 CONCLUSIONS AND DISCUSSION

Utilizing our tool, a trader will test for cointegration only "true" potential pairs, rather than running tests for all possible pairs (in a scenario of 5,000 stocks one would have $\approx$ 12 million pairs and there are close to 20,000 publicly traded stocks in the U.S.). Moreover, as mentioned above, the tool eliminates any required industry/ market knowledge required in the ruled-based testing approach. We believe it is fair to assume that using the traditional current methodologies to find the cointegrated pairs mentioned above will only be possible if cointegration tests would be performed on all possible pairs, which is inefficient and computationally "expensive".

It is important to note that stock prices are affected by a virtually infinite number of factors. From macroeconomics and market factors to geopolitical, natural disasters and entity specific factors. Users should develop and evaluate their trading strategy and consider the timeframe they test for cointegration (e.g. past 6, 12, 18,... months). Also, while the tool will help traders more efficiently find cointegrated pairs, it does not explain why, for example, GLPI is cointegrated with AY. In such situations, further research is recommended.

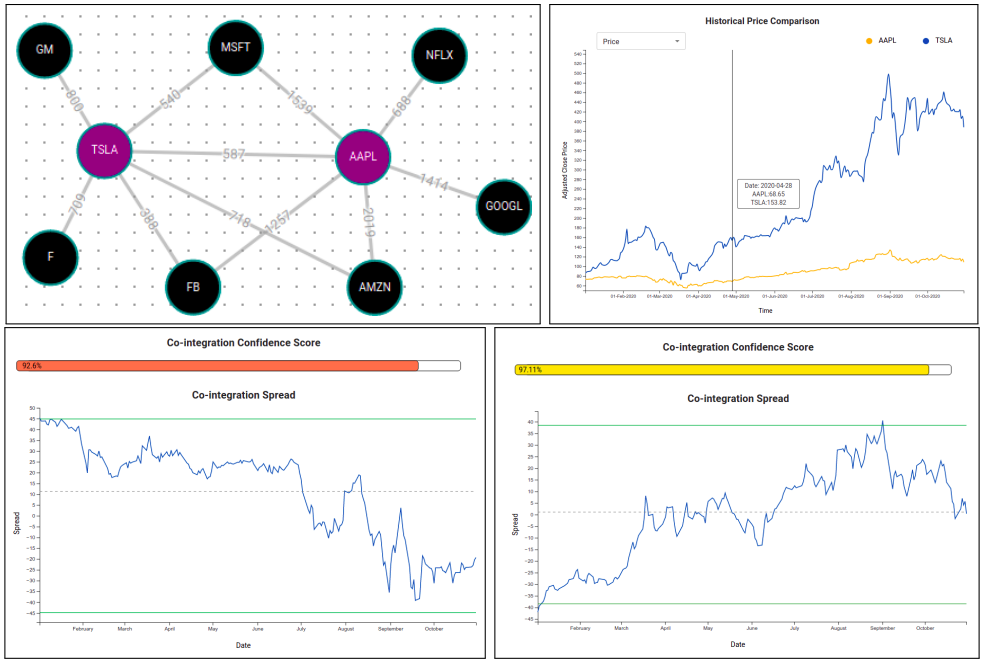

Figure 4—*(Top-Left) Network graph; (Top-Right)* ***AAPL-TSLA*** *Price plot; (Bottom-Left)* ***AAPL-TSLA*** *Co-integration; (Bottom-Right)* ***AAPL-GM*** *Co-integration*

7 TEAM WORK DISTRIBUTION

All members **contributed equally** in all the phases - project idea, literature survey, proposal and the execution phase.

Component	Owners	Tools	Tasks/Activities
Data (Market & News) Collection, Parsing & Cleaning, Processing, Storage	nbadri3, reshed6, pbaikani3, sv67	YFinance, newsfilter.io, Python, SQLite	✓ Data Collection for all the US publicly listed companies. ✓ Processing and storage of news data in SQL database for a subset of companies. ✓ Collection of news data >6 months with the same pipeline. ✓ Process and add to the database.
Core Algorithm: Tickers Network, Buckets of assets, Co-movement indices	smeduri7, reshed6	Python, statsmodel	✓ POC for constructing the network graph. ✓ Cointegration tests for stock pairs. ✓ Integrate with processed data. ✓ Define input-output model for the abstraction of core components.
Visualization: UI design, graph visuals, Interactiveness, Co-movement indices	sv67, nbadri3, pnarula6	React JS, D3, Whimsical	✓ Render basic network graph with test data. ✓ Complete network graph with processed data. ✓ Section for stock information for selected nodes. ✓ Visualization of stock data, spread, metrics. ✓ Enriching UI for user interactions.
Web Service	smeduri7, pnarula6	Python, Flask	✓ Decided Server framework and API to be exposed. ✓ Placeholder API structure created to return static data. ✓ Integrate with the core algorithm layer. ✓ Build full webservice to cater to webpages and API calls.
Cloud Infrastructure	sv67, pbaikani3, smeduri7	AWS - EC2, S3	✓ Setup data collection and processing pipelines on AWS. ✓ Deployment of the web service to power up the frontend visualization app & APIs.
Project objective	ALL	-	✓ Define Experiments and evaluation methods. ✓ Experiments tp validate the project objectives. ✓ Collect metrics to prove the hypotheses.

8 REFERENCES

[1] Engelberg, Joseph, Gao, Pengjie, and Jagannathan, Ravi (2008). "An Anatomy of Pairs Trading: The Role of Idiosyncratic News, Common Information and Liquidity". In: *SSRN Electronic Journal*. DOI: 10.2139/ssrn.1330689. (Visited on 01/14/2020).

[2] Gatev, Evan, Goetzmann, William N., and Rouwenhorst, K. Geert (2006). "Pairs Trading: Performance of a Relative-Value Arbitrage Rule". In: *Review of Financial Studies* 19, pp. 797–827. DOI: 10.1093/rfs/hhj020. URL: http://www-stat.wharton.upenn.edu/~steele/Courses/434/434Context/PairsTrading/PairsTradingGGR.pdf (visited on 04/21/2019).

[3] George, Susan and Changat, Manoj (July 2017). *Network approach for stock market data mining and portfolio analysis*. IEEE Xplore. DOI: 10.1109/NETACT.2017.8076775. URL: https://ieeexplore.ieee.org/document/8076775 (visited on 10/08/2020).

[4] Gulati, Chandra, Lin, Yan-Xia, and McCrae, Michael (Jan. 2006). "Loss protection in pairs trading through minimum profit bounds: a cointegration approach". In: *Faculty of Informatics - Papers (Archive)*, Art. No. 73803. DOI: 10.1155/JAMDS/2006/73803. URL: https://ro.uow.edu.au/infopapers/1546/.

[5] Horng, Duen, Chau, Aniket, Kittur, Jason, Hong, Christos, and Faloutsos (2020). *Apolo: Making Sense of Large Network Data by Combining Rich User Interaction and Machine Learning*. URL: https://poloclub.github.io/polochau/papers/11-chi-apolo.pdf (visited on 10/08/2020).

[6] Huck, Nicolas and Afawubo, Komivi (Nov. 2014). "Pairs trading and selection methods: is cointegration superior?" In: *Applied Economics* 47, pp. 599–613. DOI: 10.1080/00036846.2014.975417. (Visited on 08/17/2019).

[7] Jansen, S (2020). *Time-Series Models for Volatility Forecasts and Statistical Arbitrage - Machine Learning for Algorithmic Trading - Second Edition [Book]*. URL: https://learning.oreilly.com/library/view/machine-learning-for/9781839217715/Text/Chapter_9.xhtml#_idParaDest-353. (visited on 10/08/2020).

[8] Johnson, Barry (2010). *Algorithmic trading & DMA : an introduction to direct access trading strategies /*. 4Myeloma Press, URL: https://searchworks.stanford.edu/view/9155503 (visited on 10/08/2020).

[9] Krauss, Christopher, Do, Xuan Anh, and Huck, Nicolas (June 2017). "Deep neural networks, gradient-boosted trees, random forests: Statistical arbitrage on the S&P 500". In: *European Journal of Operational Research* 259, pp. 689–702. DOI: 10.1016/j.ejor.2016.10.031. (Visited on 09/13/2019).

[10] ÖZGÜR, ARZUCAN, CETIN, BURAK, and BINGOL, HALUK (May 2008). "CO-OCCURRENCE NETWORK OF REUTERS NEWS". In: *International Journal of Modern Physics C* 19, pp. 689–702. DOI: 10.1142/s0129183108012431. (Visited on 02/05/2020).

[11] Rad, Hossein, Low, Rand Kwong Yew, and Faff, Robert (Apr. 2016). "The profitability of pairs trading strategies: distance, cointegration and copula methods". In: *Quantitative Finance* 16, pp. 1541–1558. DOI: 10.1080/14697688.2016.1164337. URL: https://randlow.github.io/2016_QF_PairsTrading.pdf (visited on 08/17/2019).

[12] Rea, Alethea and Rea, William (Apr. 2014). "Visualization of a stock market correlation matrix". In: *Physica A: Statistical Mechanics and its Applications* 400, pp. 109–123. DOI: 10.1016/j.physa.2014.01.017. URL: https://www.sciencedirect.com/science/article/pii/S0378437114000211 (visited on 10/08/2020).

[13] Shimul, Shafiun, Abdullah, S, and Siddiqua, Salina (2009). "AN EXAMINATION OF FDI AND GROWTH NEXUS IN BANGLADESH: ENGLE GRANGER AND BOUND TESTING COINTE-

GRATION APPROACH". In: *BRAC University Journal* 1, pp. 69–76. URL: http://dspace.bracu.ac.bd/bitstream/handle/10361/455/Shafiu.Nahin.Shimul.pdf..?sequence=1 (visited on 10/30/2020).

[14] Skerman, Robert and Della Maggiora, Daniel (2009). "Johansen Cointegration Analysis of American and European Stock Market Indices: An Empirical Study". In: *lup.lub.lu.se*. URL: https://lup.lub.lu.se/student-papers/search/publication/1437434.

[15] Smith, R. Todd and Xu, Xun (2017). "A good pair: alternative pairs-trading strategies". In: *Financial Markets and Portfolio Management* 31, pp. 1–26. URL: https://ideas.repec.org/a/kap/fmktpm/v31y2017i1d10.1007_s11408-016-0280-x.html (visited on 10/08/2020).

[16] Taeho Jo, Duke (Jan. 2019). *Validation of Graph Based K Nearest Neighbor for Summarizing News Articles*. IEEE Xplore. DOI: 10.1109/ICGHIT.2019.00022. URL: https://ieeexplore.ieee.org/document/8866917 (visited on 10/08/2020).

[17] Vidyamurthy (2004). *CHAPTER 6 - Pairs Selection in Equity Markets - Pairs Trading: Quantitative Methods and Analysis [Book]*. URL: https://learning.oreilly.com/library/view/pairs-trading-quantitative/9781118045701/vidy_9781118045701_oeb_c06_r1.html (visited on 10/08/2020).

[18] Wen, Danyan, Ma, Chaoqun, Wang, Gang-Jin, and Wang, Senzhang (Sept. 2018). "Investigating the features of pairs trading strategy: A network perspective on the Chinese stock market". In: *Physica A: Statistical Mechanics and its Applications* 505, pp. 903–918. DOI: 10.1016/j.physa.2018.04.021. URL: https://www.sciencedirect.com/science/article/abs/pii/S0378437118304503 (visited on 10/08/2020).

9 APPENDIX

9.1 Acronyms

- AIC - Akaike Information Criterion
- BIC - Bayesian Information Criterion
- ADF - Augmented Dickey Fuller Test
- API - Application Programming Interface
- RDB - Relational Database
- POC - Proof of Concept
- AWS - Amazon Web Services
- SQL - Structured Query Language
- VECM - Vector Error Correction Model
- VAR - Vector Autoregressive Model