Department of Computer Science

Submitted in part fulfilment for the degree of BEng.

Identifying Cointegrating Relationships Between Stocks using Semantic Web Ontologies

Neel Chotai

21st May 2020

Supervisor: Dimitar Kazakov

Contents

List of Figures

List of Tables

Executive Summary

This report investigates if pairs of publicly traded companies that have links in a semantic graph are more likely to have a cointegrating relationship. We identify thousands of pairs of companies with such links stored in a semantic web ontology made up of reports submitted to the United States Securities and Exchange Commission (SEC). The ontology contains information about people, companies, and the relationship between the two, such as job title, percentage ownership, and directorship status. Using these attributes, we measured the relationship between pairs of companies and tested whether or not the stock prices were cointegrated.

This project aims to distill real-world links between companies into something quantifiable. Being able to accurately gauge the closeness of two companies can be useful for portfolio diversification, pairs trading, and other statistical arbitrage strategies. Firmly placing prominent employees in the spotlight has consequences for existing or potential future investments; a large shift in the number of employees two companies share could be used as a proxy measure to re-evaluate an investment.

The motivation for this project is simple: the nature of financial strategies leads to proprietary research and methods. We seek to change this by providing an automated way of selecting cointegrated pairs for statistical arbitrage methods. Ontologies are seeing increasing use in the fields of medicine and artificial intelligence. We seek to introduce another use case for this complex and powerful data structure as an alternative to traditional methods, specifically to the field of finance.

Our approach is centred around cointegration: the measure we used to determine whether or not companies are linked. Cointegration is often used in pairs trading to determine how close the relationship between two companies is. We sought to identify cointegrated companies with high confidence, using multiple tests to ensure that the relationship we inferred was indeed robust.

The first method tests the idea that pairs of companies that share a greater number of employees are more likely to be cointegrated than those selected at random. Therefore, by creating a list with pairs of companies and analysing the number of relationships they have, we can create a subset of companies with the most connections in this dataset and test our hypothesis.

The second method uses sampling to randomly select companies, providing a better picture of the entire dataset. Two test sets were created, one containing pairs of companies that shared employees and one containing pairs of companies that shared directors - both given by the ontology. Iteratively, the constraint on the minimum number of relationships between the pairs each set was increased to identify the relationship between sharing a greater number of attributes and the presence of cointegration. A control set was created from randomly generated companies, from the Nasdaq stock exchange, to assist in this comparison.

The investigation determined that the hypothesis has some merit. In both tests, pairs selected by the number of shared employees substantially outperformed the control set. A point of contention was the directorship set which produced no measurable increase in selecting cointegrated companies compared to random generation. Future work could concentrate more specifically on the number of employees two companies share, which was the most promising method. More data would be another useful addition - comparing the cointegration of companies over time, especially in the case both companies lose or gain mutual employees, would provide a meaningful insight into the long-term implications of this relationship.

There are no legal, social, ethical, professional, and commercial issues that arise as a consequence of this project.

1 Introduction

Pairs trading is a form of statistical arbitrage where the trader takes out both a short and long position simultaneously. Highly correlated stocks that cointegrate well are the best candidates for pairs trading as the price tends to move together over time. Murray describes cointegration humourously using the example of a drunk and her dog [1]. Observing the pattern of a drunk after a night of drinking would look completely random, much like the path of a wandering dog - the dog will meander to wherever its nose takes it. If, however, the dog belongs to the drunk and the drunk calls for the dog, the dog will remain close to the drunk - hence the two wander aimlessly, together. This is precisely the concept of cointegration and what we hope to capitalise on.

Information from SEC reports stored in web ontologies, such as directorship, percentage ownership, and company affiliation, may be key to finding highly cointegrated pairs. For example, the case may be that two companies that share directors have a high degree of correlation, cointegrate well, and so are ideal candidates for pairs trading. Using this information and historical stock data, we can test our hypothesis.

This report marries together the fields of computer science and economics by seeking to identify whether attributes shared between publicly traded companies have a statistically significant effect on their pairs trading performance. Given a pair of companies that share directors or other employees, we expect the ratio of their prices (i.e., the spread) to remain constant over time. On occasion, there will be a divergence in the spread which may be caused by important news, liquidity, or volatility. Over time, we expect that this spread will return to normal and thus we can profit from this by making a pairs trade.

We believe that companies with a greater number of relationships within our web ontology, populated with information from SEC reports, will make superior candidates for pairs trading. Our methodology will test to see if companies that share these attributes cointegrate well, visually display their relationship, and show how the stocks of these companies move over time.

2 Literature Review

This chapter will assess the viability of this project by focusing on three main subject areas. We will be looking at the usage of ontologies, the performance of pairs trading, and methods for obtaining pairs to decide whether or not cointegration is the optimal strategy.

2.1 Ontologies

The Semantic Web is an extension of the World Wide Web with the goal of making stored data machine-readable. In contrast to the World Wide Web, which links together data machines have no understanding or context of, the Semantic Web provides an account of meaning, a logical connection [2]. In the World Wide Web, two databases may use completely different identifiers for the same concept, a program intending to utilise this data needs to know that these two terms mean the same thing. Bernes-Lee et al. describe ontologies as the solution for this assortment of uncontextulised data using the idea of subjects and relations to describe links between items of data, inference rules to allow for more effective ways to manipulate data, and a unified interface to increase software efficiency [3].

Ontologies are a representation of concepts within a domain, the concepts being categories, properties, and relationships between data and entities. Simply put, an ontology shows us characteristics of a subject and how these are related to other characteristics and subjects. In our case, the subjects will be people and companies; characteristics will be things like job title, directorship status, and the stock ticker. Ontologies have multiple benefits, as have been noted by Uschold and Gruninger [4], such as ease of programming implementation, informative visualisation but the principal among them being that ontologies remove the barrier caused by disparate backgrounds, languages, tools, and techniques by providing a single unified interface to a dataset. Ontologies also provide more knowledge about the domain of the data, allowing us to more accurately and easily model relationships.

Ontologies have been used in other fields for much the same reason we believe they will be a better source of information than a simple database. Frank examines how ontologies can be used to improve the consistency of data coming from many different sources [5]. Instead of consistency constraints and database schemas, ontologies easily integrate data from different sources into a single system, eliminating erroneous data and allowing our dataset to become more extensible, giving us an easy way to new add data in the future. Smith et al. have described the use of ontologies in the biomedical field for the purpose of disambiguation [6]. Smith et al. found that ontologies are more consistent and unambiguous compared to a traditional database in the case of medical vocabulary where two completely unrelated procedures or medicines may share the same name or abbreviation. This is particularly useful for us - for example, if multiple people have the same name, ontologies provide a better way to link together information and an easy interface to visualise these relationships.

2.2 Pairs trading

Pairs trading is a market-neutral trading strategy that allows traders to profit in virtually any market conditions. The strategy is a form of statistical arbitrage, a class of short-term financial strategies that are highly quantitative in nature and employ mean reversion models. Pairs trading, as the name suggests, consists of buying two positions: one asset is bought long and a similar asset is sold short for a higher price [7]. The assumption being that, by the time the assets are delivered, the prices will be close to equal allowing the trader to profit by the amount the assets initially diverged by - this is known as a convergence trade.

Pairs trades have shown to be profitable by simply choosing two stocks that have moved together in the past [8]. The logic behind this strategy is straightforward: if a pair of stocks has historically moved together and one has diverged, we can short the underperforming stock and long the outperforming stock. Assuming that history repeats itself and the stocks converge, the trader can profit from a relatively low-risk trade.

In a study by Huck and Afawubo comparing multiple ways to select pairs, cointegration was found to be the superior method exhibiting high and robust positive alpha [9].

2.3 Cointegeration

Two time series variables are cointegrated if the expected value of the ratio converges to the mean over time. Granger defines cointegration as follows [10]: Take two series, x_t and y_t, each of which is integrated of order 1 ($I(1)$) and has no drift or trend in mean. Generally, any linear combination of these series is also $I(1)$. However, if there exists some constant A, such that

$$z_t = x_t - Ay_t$$

is $I(0)$, x_t and y_t are said to be cointegrated with A being the cointegrating parameter. Several testing methodologies are available to find out if two time series are cointegrated, the most popular being the Engle-Granger and Johansen tests.

The Engle-Granger methodology is carried out in two steps. The first step is to create residuals based on the static regression of two time series and the second step is to test these residuals for the presence of unit roots, typically done using the augmented Dickey-Fuller test [11] [12]. If the p-value returned by this test is sufficiently low, we can reject the null hypothesis of no cointegration.

The Johansen test can test for cointegrating relationships between more than two time series (eliminating the issue of choosing a dependent variable) and allows for more than one cointegrating relationship, unlike Engle-Granger [13]. This methodology is subject to asymptotic properties and requires a large sample size or unreliable results are produced. Two results are returned by the Johansen test: the trace statistic and the maximum eigenvalue statistic. The null hypothesis for both tests is that the number of cointegrating relationships (given by linear combinations) is at least one. For the trace test, the null hypothesis is given by $H_0 : K > K_0$ as opposed to the alternate hypothesis $H_a : K = K_0$. K_0 is set to zero to test if the null hypothesis can be rejected and, if it is, we deduce there is a cointegrating relationship in the sample. For the maximum eigenvalue test, the alternate hypothesis is given by $H_a : K_0 + 1$. The difference between the two methods is that the maximum eigenvalues test gives an estimator of the number of linear combinations: if $K = K_0$ and the null hypothesis is rejected, we can deduce there is one possible outcome that produces a stationary process. However, if $K_0 = N - 1$ and the null hypothesis is rejected, there are N possible linear combinations.

Both methods are far from perfect, for example, research by Do et al. compares Engle and Granger's 2-step approach with Johansen's, finding the former to be influenced by the ordering of the variables and, on occasion, returning spurious estimators [14]. However, research carried out by Gonzalo and Lee shows that, for most situations, Engle-Granger is more robust than Johansen's likelihood ratio test [15]. Gonzalo and Lee recommend using both Engle-Granger and Johansen tests in order to detect the possibility of a pitfall and subsequently increase the chance of avoiding it.

Cointegration is a commonly used measure to show a statistically significant connection between two time series. We will be focusing on individual securities as time-series but research from Khan has shown cointegration can be used on a wider scale by examining the convergence of global markets [16]. This paper displays the interconnectedness of global stock markets, a phenomenon that has been consistently observed [17] [18] [19], lending credence to the idea that cointegration is a good measure of connection.

The reason cointegration is preferable for pair selection instead of correlation is uncertainty. Correlation indicates that two securities are linked but says nothing about their magnitudes. For example, two stocks may climb together but at vastly different rates. Cointegration removes this uncertainty by identifying pairs that do not drift too far away from each other over the long term.

3 Problem Analysis

3.1 Objectives

Our main intention is to elucidate whether these links between companies have any effect on their market price. This information could provide a substantial benefit to a pairs trading strategy which ordinarily requires expert knowledge on position sizing, market timing, and good decision making. Automating the decision making factor could lead to a more profitable and reliable strategy.

Pairs trading is not the only use for identifying closely correlated pairs: cointegrated pairs could be used, for example, to diversify investment in a sector. If an investor felt bullish on semiconductors, they may want to take out multiple positions in cointegrated companies instead of a single position in one.

An ancillary objective is to call attention to the fact that companies are made up of people. Should we find statistically significant evidence that two companies linked by one or several people are cointegrated and typically move together, this has interesting consequences on future investments and may shift over the perspective of looking at a company to looking at pioneering individuals.

3.2 Motivation

Huck and Afawubo remark that strategies implemented are proprietary in nature [9], which is one of the main reasons this project exists. The process of deriving potentially relevant financial information from a web ontology lends itself to statistical arbitrage strategies that are not limited to pairs trading. Time series analysis, autoregression, and pattern finding techniques can all utilise cointegrated company selection. Automated algorithmic strategies are not the only avenue to utilise this research however, the results provided may be a useful indicator to an investor for a current or potential future investment or could be used as one component of a more complex strategy/model.

Another reason is to provide some additional information to the efficient market hypothesis [20], a theory that has led to much discussion and dispute. The information garnered from web ontologies may provide credence to classifying our market as semi-strong form.

Understanding market trends is useful for more than merely turning a profit. Market analysis can aid us in extrapolating information about sector saturation, the emergence of new products and technologies, and understanding of people's needs through demand.

3.2.1 Alphabet

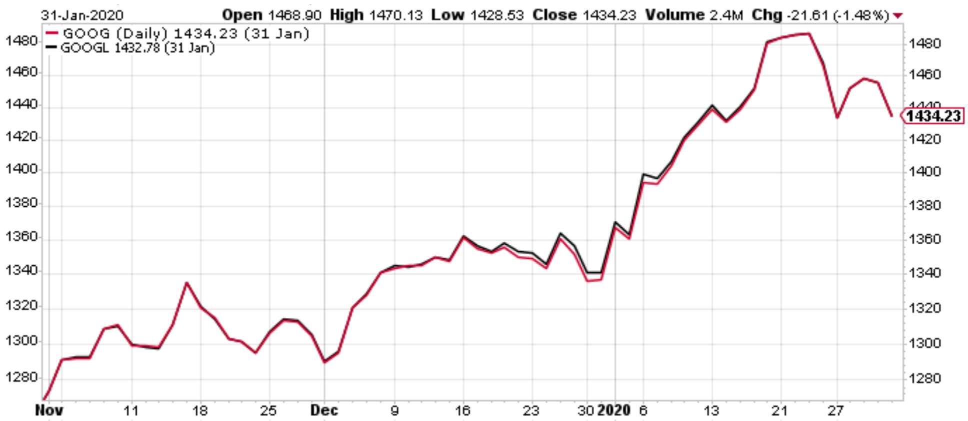


Figure 3.1: Stock price of GOOG (Alphabet Inc - Class C) and GOOGL (Alphabet Inc - Class A) from 2019/11/1 to 2020/1/31.

Alphabet is traded as both GOOG (class C) and GOOGL (class A). Running the Engle-Granger cointegration test gives a p-value of 0.0022, indicating that the two pairs are cointegrated with 99% confidence. The most likely explanation for this relationship is that the two stocks belong to the same company.

3.2.2 Visa and Mastercard

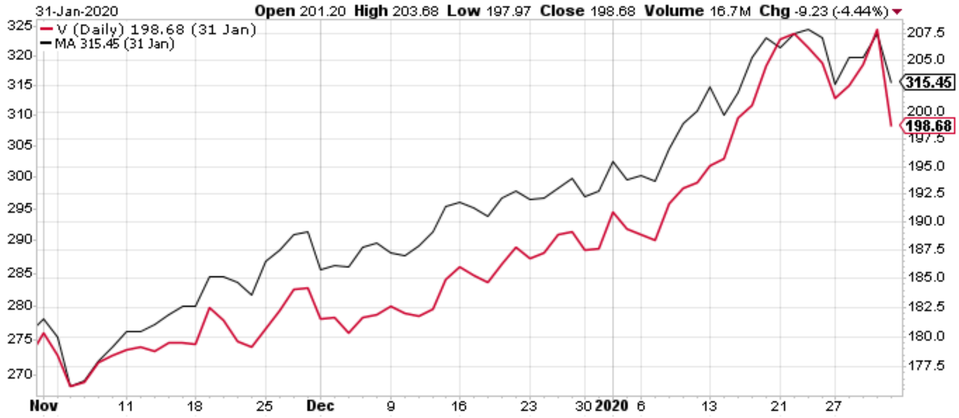


Figure 3.2: Stock price of V (Visa Inc - Class A) and MA (Mastercard Inc - Class A) from 2019/11/1 to 2020/1/31.

The stock prices of Visa and Mastercard tend to move together based on historical information. The Engle-Granger cointegration test returns a p-value of 0.0008, indicating that the pairs are significantly cointegrated. Knowing about this relationship gives a little more information about the companies through purely quantitative analysis, we can infer that the companies may be in the same sector, may be competitors or may share significant people.

3.3 Problems

The automatic identification of cointegrated companies based on attributes is not without fault. Our approach of using SEC reports may lead to operating on outdated information; publicly traded companies are only required to submit reports, at minimum, quarterly. Given that we intend to model long-term relationships, this becomes less of a concern but this risk must nonetheless be kept in mind.

Pairs trading is one of the financial fields that utilises cointegration, however the strategy is not without risk. When two securities begin to drift apart (i.e., do not revert to the mean), strict risk management rules or manual intervention may be required to mitigate losses.

3.4 Hypothesis

Our hypothesis is based on the fact that that cointegration is a measure of some statistically significant connection between two securities and we want to detect companies that are cointegrated with a high probability. We posit that selecting pairs of companies based on specific relationships or links, be it sector, shareholders, or employees, in the semantic graph, could increase the probability of identifying pairs of companies with desirable properties.

We hypothesise companies that share directors or a large amount of other prominent employees are more likely to be cointegrated than companies selected without this constraint, and that there is a quantifiable difference between companies that are cointegrated and companies that are not, which can be demonstrated using web ontologies.

4 Design and Implementation

4.1 Methodology overview

The method to test our hypothesis is split into two parts: selecting pairs and testing cointegration.

In order to select pairs, we will be utilising information garnered from SEC reports, in the form of a web ontology, and the Nasdaq stock exchange to compile three sets of pairs of companies. One set of pairs will be selected at random, one set of pairs will be selected by shared directorship and one set of pairs will be selected by shared prominent employees. We believe that the latter two sets have a higher likelihood of containing cointegrated pairs.

To test for cointegration, we will be using historical stock data for the companies from the Nasdaq stock exchange. Each pair in the three sets will be tested to see how well they cointegrate and the total number of cointegrated pairs will be compared across sets.

As an additional test, we will randomly sample each of the sets and test how many pairs are cointegrated, comparing the averages across sets with the intent of limiting bias as much as possible.

4.2 Data

The data from the web ontology is stored in N-Triples format and gives different information depending on the entity. The ontology file consists exclusively of lines containing either statements or comments. N-Triples statements have four parts: the subject, predicate, object, and a full stop which marks the end of the statement. N-Triples comments consist of the information pertaining to the predicate.

```
<http://sec.com/0001614838> <http://xmlns.com/foaf/0.1/name> "Neeleman
    Stephen"^^<http://www.w3.org/2001/XMLSchema#string> .
<http://sec.com/0001614838> <http://schema.org/jobTitle> "Founder and Vice
    Chairman"^^<http://www.w3.org/2001/XMLSchema#string> .
<http://sec.com/0001614838> <http://york.ac.uk/cik> "0001614838"^^<http://
    www.w3.org/2001/XMLSchema#string> .
<http://sec.com/0001614838> <http://york.ac.uk/isdirector> "true"^^<http://www.
    w3.org/2001/XMLSchema#boolean> .
<http://sec.com/0001614838> <http://york.ac.uk/isofficer> "true"^^<http://www.
    w3.org/2001/XMLSchema#boolean> .
<http://sec.com/0001614838> <http://york.ac.uk/is10percentowner> "false"^^<
    http://www.w3.org/2001/XMLSchema#boolean> .
<http://sec.com/0001614838> <http://york.ac.uk/isother> "false"^^<http://www.
    w3.org/2001/XMLSchema#boolean> .
```

Figure 4.1: Example of information stored for an entity of type Person.

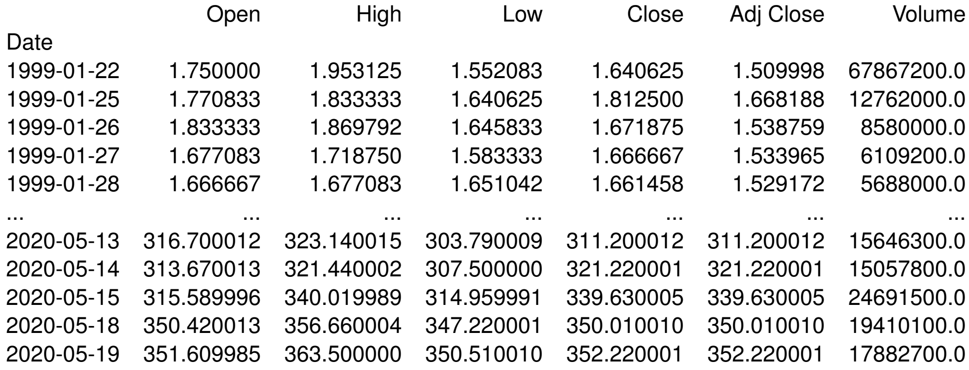
The predicates in figure 4.1 indicate the name and job title of the entity along with information about the position of the person within the company: whether they are a director, officer, shareholder, or none of these things but still a prominent person affiliated with the company. The CIK predicate indicates the Central Index Key of the person within the SEC database.

```
<http://sec.com/0000320193> <http://york.ac.uk/cik> "0000320193"^^<http://
    www.w3.org/2001/XMLSchema#string> .
<http://sec.com/0000320193> <http://xmlns.com/foaf/0.1/name> "APPLE INC
    "^^<http://www.w3.org/2001/XMLSchema#string> .
<http://sec.com/0000320193> <http://york.ac.uk/tradingsymbol> "AAPL"^^<http
    ://www.w3.org/2001/XMLSchema#string> .
<http://sec.com/2017/QTR1/1258883/4/0001628280−17−001483.txt> <http://
    york.ac.uk/periodreport> "2017−02−16"^^<http://www.w3.org/2001/
    XMLSchema#string> .
<http://sec.com/2017/QTR1/1258883/4/0001628280−17−001483.txt> <http://
    york.ac.uk/reporttype> "4"^^<http://www.w3.org/2001/XMLSchema#string> .
```

Figure 4.2: Example of information stored for an entity of type Company.

The predicates in figure 4.2 indicate the name and ticker symbols of the company along with the date of the relevant SEC report and the type of form. In this instance, the comment for the <http://york.ac.uk/reporttype> predicate shows that this information was given by a statement of changes in beneficial ownership of securities form. The CIK predicate indicates the Central Index Key of the company within the SEC database.

The historical stock data is provided by the yfinance Python module (through Yahoo Finance) [21] and comprises of information for the stock using the date as an index. For example, a query on NVDA provides the information displayed in figure 4.3.



```
                 Open        High         Low       Close   Adj Close      Volume
Date
1999-01-22   1.750000    1.953125    1.552083    1.640625    1.509998  67867200.0
1999-01-25   1.770833    1.833333    1.640625    1.812500    1.668188  12762000.0
1999-01-26   1.833333    1.869792    1.645833    1.671875    1.538759   8580000.0
1999-01-27   1.677083    1.718750    1.583333    1.666667    1.533965   6109200.0
1999-01-28   1.666667    1.677083    1.651042    1.661458    1.529172   5688000.0
...               ...         ...         ...         ...         ...         ...
2020-05-13 316.700012  323.140015  303.790009  311.200012  311.200012  15646300.0
2020-05-14 313.670013  321.440002  307.500000  321.220001  321.220001  15057800.0
2020-05-15 315.589996  340.019989  314.959991  339.630005  339.630005  24691500.0
2020-05-18 350.420013  356.660004  347.220001  350.010010  350.010010  19410100.0
2020-05-19 351.609985  363.500000  350.510010  352.220001  352.220001  17882700.0
```

Figure 4.3: Information provided by yfinance on NVDA (NVIDIA Corp).

The field we are interested in is the open price which is what we will be using to test for cointegration. In total, we will be using two years of stock data (between 2016-2018 inclusive) to provide an accurate image of the relationship between two stocks. To eliminate any missing data, i.e., if one ticker has information for a date that the other ticker doesn't, we merge both lists of time series by date and drop any rows with NaN (missing) values.

4.3 Pair selection

There are two methods we will be using to choose company pairs: companies that share at least one attribute (employees) and companies that share directors. Both methods will require using SPARQL to query the ontology.

Selecting companies will be done by selecting a Person entity that works at a company and works at another company, provided by the <http://york.ac.uk/worksat> predicate with the constraint that both companies cannot be the same.

```
SELECT ?person ?p ?t1 ?t2
WHERE { {
    ?person <http://york.ac.uk/worksat> ?company .
    ?person <http://york.ac.uk/worksat> ?othercompany .
    ?person <http://xmlns.com/foaf/0.1/name> ?p .
    ?company <http://york.ac.uk/tradingsymbol> ?t1 .
    ?othercompany <http://york.ac.uk/tradingsymbol> ?t2 .
    FILTER(?t1 != ?t2)
} }
```

Figure 4.4: SPARQL query to extract employee information.

Selecting companies that share a director differs in exactly one regard: the constraint that the person we select must be a director, provided by the <http://york.ac.uk/isdirector> predicate.

```
SELECT ?person ?p ?t1 ?t2
WHERE { {
    ?person <http://york.ac.uk/worksat> ?company .
    ?person <http://york.ac.uk/isdirector> true .
    ?person <http://york.ac.uk/worksat> ?othercompany .
    ?person <http://xmlns.com/foaf/0.1/name> ?p .
    ?company <http://york.ac.uk/tradingsymbol> ?t1 .
    ?othercompany <http://york.ac.uk/tradingsymbol> ?t2 .
    FILTER(?t1 != ?t2)
} }
```

Figure 4.5: SPARQL query to extract directorship information.

Both queries return the relevant SEC report, the name of the person, and the tickers for both companies the person is affiliated with. This information gets fed into a dictionary where the key is the pair of companies and the value is a list containing the people affiliated with the company. Once the dictionary is fully populated, the values for each key will be processed to remove duplicates after which, the list of names for each key will be replaced with the total amount of people (length of the array).

4.4 Testing cointegration

Our objective is to model long-term relationships between companies. In order to test this hypothesis, we will be utilising both the Engle-Granger two-step test and the Johansen test, both provided by the statsmodel Python package [22] [23].

For the Engle-Granger test, the values returned are the t-statistic of unit-root test on residuals, MacKinnon's approximate asymptotic p-value [24] and critical values for the test statistic at the 1%, 5%, and 10% levels based on the regression curve. The p-value is the value we are most interested in: a p-value less than 0.05 allows us to reject the null hypothesis and indicates that the two time series are cointegrated.

For the Johansen test, the values returned are the trace and maximum eigenvalue statistics. The value we will be using to reject the null hypothesis is the trace statistic, the reason for choosing this over the maximum eigenvalue statistic is because we are only concerned with the presence of cointegration and not the number of linear combinations. Preliminary testing has revealed that the Johansen test is stricter than Engle-Granger in accepting pairs; as such, we will be using the trace statistic to reject the null hypothesis at a 95% confidence level to test whether pairs are cointegrated.

4.4.1 Method 1: total cointegrated pairs

The ontology consists of information contained within SEC reports released in Q1 of 2017. As such, pairs will be tested for cointegration using historical stock data between Q3 2016 - Q2 2017. If both tests indicate a cointegrating relationship between this time frame, the pair will be added to a set of valid pairs - the amount of pairs in this set allows us to evaluate our hypothesis.

In order to check for spurious relations, we will test different sets of data. Our control set will be tickers selected at random from actively traded symbols on the Nasdaq's stock exchange [25]. Our directors set will be made up of companies that share at least one director and our employees set will be made up of companies that share at least one employee. The directors and employees sets will be selected from the ontology.

Both the directors set and employees set will be sorted by descending number of shared attributes so that pairs with the highest number of shared attributes will be selected first. Pairs of companies in each set will be tested for cointegration to see if there is a difference between companies selected at random and companies we hypothesise will be cointegrated by virtue of sharing attributes.

All three sets will be preprocessed:

- Any reflexive pairs (e.g. ("TSLA", "TSLA")) will be removed. This constraint is in both SPARQL queries but will be included in Python to ensure consistency.
- If there is an instance of a pair and reversed pair (e.g. ("INTC", "AMD") and ("AMD", "INTC")) in a set, one will be removed. In order

to choose the pair to be removed, both pairs will be tested for a cointegrating relationship. If any such relationship is found in exactly one pair, that pair will be retained. If a cointegrating relationship or no cointegrating relationship is found in both pairs, the pair with a lower index will be retained.

- Any stock that has less than 251 days of stock history will be removed, an accurate measurement of cointegration cannot be garnered from a stock with this little data. This number is given by the total number of trading days between Q3 2016 and Q2 2017.
- In order to eliminate any potentially decommissioned tickers, any stock that has less than 251 of stock history between Q3 2017 - Q2 2018 inclusive will also be removed.

4.4.2 Method 2: random sampling

An issue with method 1 is that there is some bias in preprocessing sets due to an affinity for cointegrated pair retention when a pair and reversed pair exist within the list. In order to offset this, testing will also be done in the form of random sampling without replacement. When sampling, all preprocessing constraints will be utilised except filtering pairs and reversed pairs, both of which will remain in the set if this situation occurs.

We will be sampling the data 10 times and, in each sample, 30 pairs will be selected at random (without replacement) from the directors and employees sets. In addition to this, we will test the directors and employees sets with the constraint that companies share at least one attribute, three attributes, and five attributes separately. The number of pairs has been limited to 30 per sample as preliminary testing shows that the number of pairs available reduces drastically the higher the attributes and limiting the pairs will limit selection bias.

The control set will also be sampled in a similar fashion: 30 random companies will be selected, per sample, from actively traded companies on the Nasdaq stock exchange. In order to compensate for testing multiple attribute counts in the directors and employees sets, the sampling process for the control set will occur 30 times to match the total generated samples in the directors and employees sets.

Each pair in each sample will be tested for cointegration with the total number of cointegrated pairs per sample being measured. The average of the total cointegrated pairs over all ten runs (30 in the case of the control set) will be measured. The results obtained for the control set will be used to compare across the entire set of results from the directors and employees sets.

5 Results and Evaluation

5.1 Method 1: total cointegrated pairs

For the control set, 150 pairs of companies were generated randomly from companies traded on the Nasdaq stock exchange, preprocessed and tested for cointegration between Q3 2016 - Q2 2017 inclusive. Of these pairs, 9 (6%) had a p-value less than 0.05 and passed the Johansen test with 95% confidence, allowing us to reject the null hypothesis of no cointegration.

For the directors set, 10186 pairs of companies were generated with one or more directors shared with another company. Pairs were selected and preprocessed from the director set until 150 companies were returned. Of these pairs, 11 (7.33%) had a p-value less than 0.05 and passed the Johansen test with 95% confidence, allowing us to reject the null hypothesis of no cointegration.

For the employees set, 16218 pairs of companies were generated with one or more employees with another company. Pairs were selected and preprocessed from the employee set until 150 companies were returned. Of these pairs, 18 (12%) had a p-value less than 0.05 and passed the Johansen test with 95% confidence, allowing us to reject the null hypothesis of no cointegration.

Pairs and cointegrated pairs generated from each set can be found in appendix A.

Directors set	
Shared attributes	Number of pairs
10+	5
9	2
8	2
7	29
6	2
5	0
4	2
3	18
2	90

Employees set	
Shared attributes	Number of pairs
10+	8
9	7
8	19
7	116

Table 5.1: Tables showing the distribution of pairs selected with shared attributes.

The total percentage of cointegrated companies shows that there is no statistically significant difference between choosing pairs at random and choosing pairs that share directors. In contrast, the employees set has a substantially greater amount of cointegrated stocks compared to both the control and directors sets.

It is also of note that there is a higher concentration of pairs with a high number of shared attributes in the employees set over the directors set.

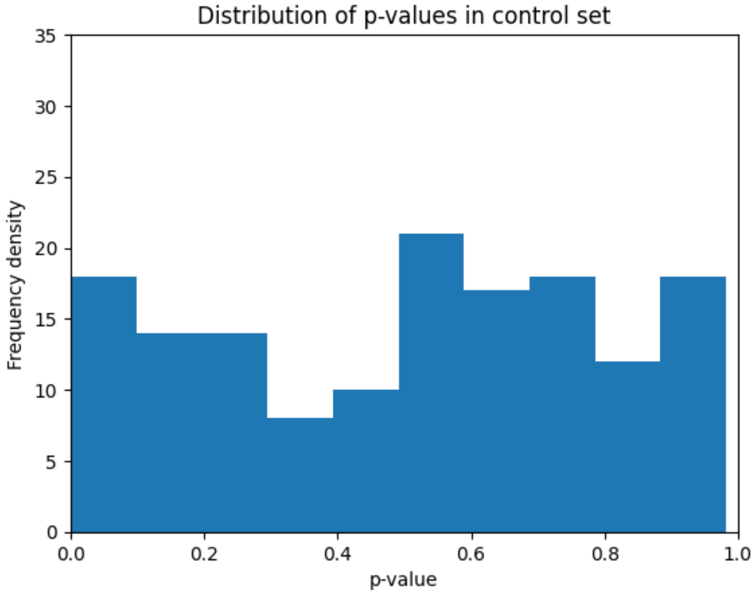


Figure 5.1: Histogram showing the distribution of p-values for every pair in the control set.

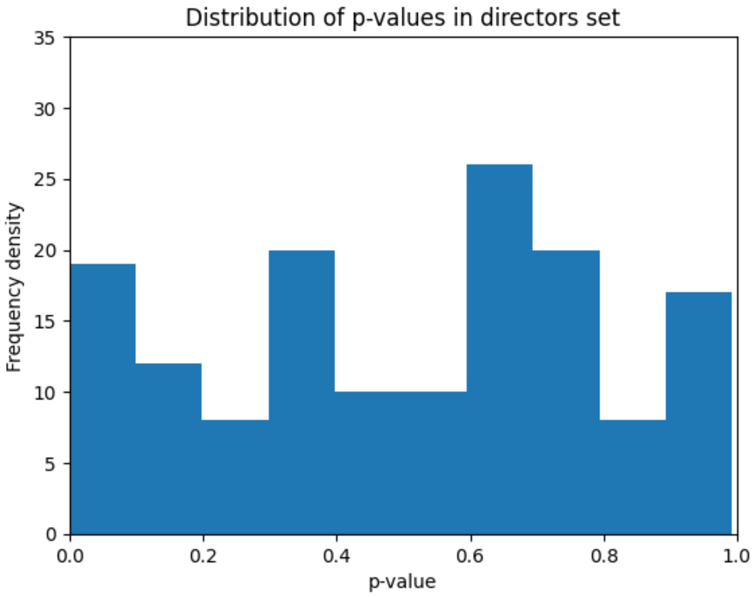


Figure 5.2: Histogram showing the distribution of p-values for every pair in the directors set.

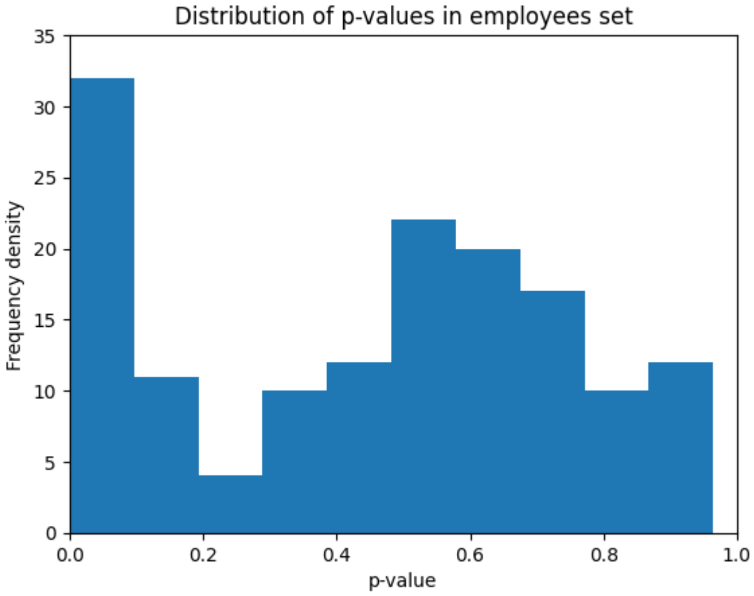


Figure 5.3: Histogram showing the distribution of p-values for every pair in the employees set.

Examining the histograms above, we can note that there is not a significant degree of difference between the p-values of pairs in the control set and the directors set; however, there is a higher frequency of p-values clustered around 0 in the employees set. The Fisher-Pearson coefficient of skewness for the control sets is given by -0.161898; the Fisher-Pearson coefficients for the directors and employees sets are given by -0.192676 and -0.193893, respectively. This indicates that all sets are approximately symmetric, however the employees and directors sets are slightly more negatively skewed compared to the control set.

5.2 Method 2: random sampling

		Number of shared attributes		
		At least 1	At least 3	At least 5
Cointegrated pairs (average)	Directors set	1.3	0.6	1.1
	Employees set	0.8	1.6	2.0
Total pairs	Directors set	3543	141	77
	Employees set	4208	483	336

Table 5.2: Table showing the average number of cointegrated pairs, for the directors and employees sets, from sampling.

After taking ten samples of 30 pairs of companies a total of 3 times, the average number of cointegrated companies in the control set was 1.5.

From the sampled data displayed in table 5.2, we can deduce that the employees set has a much higher likelihood of containing cointegrated pairs given a large number of shared attributes. It is apparent that there is a correlation between the number of shared attributes and the number of cointegrated companies in the employees set, however this is not shared by the directors set which performs worse than the control set despite restraining the attribute count.

The total cointegrated pairs per sample can be found in appendix B.

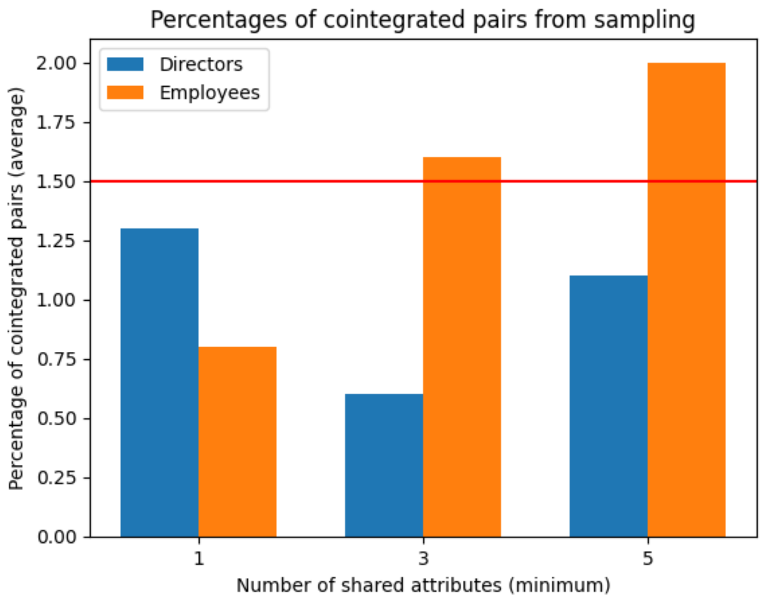


Figure 5.4: Bar graph showing average cointegrated pairs for directors and employees sets from sampling with red line to indicate control set average.

5.3 Evaluation

Overall, the automated system for selecting pairs seems to be effective in the case of companies sharing employees. There appears to be a significant link between the number of employees a pair of companies share and the likelihood of these companies being cointegrated. Both methods indicate that the directors within companies are less important than the total number of employees both companies share when it comes to the movement of their stocks.

The directors set provided disappointing results, showing no evidence that companies that share directors are more likely to be cointegrated, though this may be due to limited data. In both methods, the employee set had a much higher amount of total pairs by virtue of omitting the directorship restraint. It is possible that shared directors contribute to the cointegration of a pair of companies but we have no evidence to support this and too little data to refute it.

The employees set provided promising results in both methods, drastically outperforming the control and directors sets for selecting cointegrated companies. Both methods indicate that there is a positive correlation between the number of employees a pair of companies share and the presence of a cointegrating relationship. However, there is no evidence to suggest that companies that share a small number of employees (less than five) are likely to be cointegrated.

The inclusion of a control set shows us the difference between correlation and computation; an investor could randomly generate cointegrated companies and outperform company selection by directorship. However, there is empirical evidence of a link between the number of prominent employees two companies share and the cointegration of their stocks, which may have other implications or desirable traits over companies selected randomly. For example, a drastic drop in shared employees between quarters could signal investors that something is amiss. Similarly, a large influx of employees could lead to a closer cointegrating relationship and may cause the investor to reconsider the size of their position.

The interfaces used to obtain the data, yfinance and SPARQL, performed well; yfinance provided immediate access to accurate historical stock information and SPARQL enabled easy querying over a large dataset while avoiding the pitfalls associated with traditional databases. One issue that arose was inconsistent patterns with classes of tickers: for example, Berkshire Hathaway Inc. is traded as both BRK-A (class A) and BRK-B (class B). The way a ticker is formatted was a problem: BRK-A can take the form BRK.A, BRK/A, or BRKA depending on the source providing the stock information and similarly for BRK-B. Manual intervention was required to fix these inconsistencies.

6 Conclusion

Our hypothesis is that companies that share relationships in the semantic graph, specifically employees, are more likely to be cointegrated than companies selected without this constraint. Our results somewhat confirm this: there appears to be a link between companies that share a significant number of employees and a cointegrating relationship but no such link exists in pairs of companies that share any number of directors.

Given the information used in the report has been in the form of publicly available SEC reports and historical stock information, the results provide some credence to classifying our stock markets as semi-strong form. Semi-strong form efficiency contends that publicly available information is priced into securities, and changes to the equilibrium reflect changes to that information. Knowing that companies that share a large number of employees have stocks that tend to move together, we can infer that this information is priced into securities.

For further work, more data would have been useful in order to test cointegration over time. As the SEC reports in the ontology were exclusively from Q1 2017 and cointegration was tested for months in which we did not have SEC data, cointegration may have been spurious or potentially cointegrated companies could have slipped through. As an example, in our Motivation, we state that the companies Visa and Mastercard are cointegrated but this pair fails our cointegration test within the time period tested.

When testing cointegration, we have been testing the opening prices of the stocks for both companies. However, this does not take into account that pairs of companies may be cointegrated with the market rather than each other. In order to eliminate this, an index fund could be used as a confounding variable. For example, for companies based in the United States, the S&P 500 (SPX) index could be tested for cointegration per company using the Engle-Granger test and for the pair of companies and the index using the Johansen test.

Some items of data had erroneous formatting: for example, some ticker strings included symbols, additional unnecessary characters, and names of exchanges prefixing the ticker. Whilst most of these errors were fixed manually, incompatibilities with looking up tickers may have caused some valid items of data to be discarded. In future work, effort should be taken to ensure the input ontology is as clean as possible.

With that being said, the information provided by the ontology was instrumental in delivering the results obtained. With a slightly modified approach, the same methodology could be applied to a new ontology with different data. Research by Liu et al. has delved in to feature extraction from news articles and embedding this data in a knowledge graph [26], it may be possible to use this or similar techniques to create a real-time, more data-rich ontology.

Focusing exclusively on the number of employees two companies share could lead to better results in the future. The population of the employees set is abundant at a high number of attributes and lends itself to a more fine-grained approach to company selection. Displaying that this link exists opens the door to a closer inspection of this particular relationship.

A Method 1: sets of pairs

Control set pairs:

GBLI/EQNR, BCM/HE, TSC/FTI, XLY/DISCK, BDJ/SOYB, MDC/QAT, CLIR/DX, TCBK/CIGI, GAB/RBCAA, OPHC/MPWR, IEUR/ANH, EMFM/ENR, GRPN/DRNA, MEIP/PPH, THW/HLT, DWX/STXS, CINR/BWZ, XNTK/NKX, BIMI/AGTC, ENOR/ATHX, SEED/AVK, NSTG/XLG, PWV/IAGG, BPOPM/MITK, PRIM/GNE, CSWC/BANC, FFR/ZIXI, SCIJ/T, TLK/AAON, DPG/ICE, MNE/PY, EDC/PPX, SITC/FITB, FTS/TWO, JDST/BLDR, PRAH/DCOM, IIF/INWK, ERIE/GENC, DSE/CURE, TVC/AKER, EMKR/XRLV, FUL/CORR, PFSW/MDYV, AADR/PEZ, ETN/KIE, JKH/CZNC, PIN/FEX, TFC/UAN, PROV/STL, IJS/NWN, RY/MPWR, PWOD/EXAS, WNS/MARK, CHAD/CODI, QNST/PKX, BSET/XELB, SANM/LSXMA, AHT/DGT, CHII/PDCO, LTPZ/PEIX, BTG/HDS, TZA/KN, MGEN/HYI, BC/MYI, DCP/SCHE, LGL/PUK, MBRX/TGT, IOSP/SPXT, CTB/SMEZ, LLNW/DJCI, CFG/CBMG, SSYS/SMM, UG/ITRI, MUB/JBK, TMQ/CPE, WSC/PED, MSVB/SLQD, EPAC/RIGL, PSB/OLD, SHOP/CETX, FRBK/DBB, FLL/AGI, JJSF/SPLK, NCTY/PCG, MNST/CSF, UNM/SFHY, PKX/SIG, KKR/CECE, INS/ASRT, CCOI/NAOV, EMX/GLD, RLJ/WEX, NOVT/PME, EFX/MSEX, ATO/CLNY, ANH/ENTA, HIG/GTT, JBSS/SUM, TWMC/ZBH, HTHT/MRIN, COLB/IGV, CPHC/CASI, BHB/CL, ZGNX/PRPL, IWD/ECL, SUP/HDS, JPM/JTD, CALX/SSO, HCM/ITW, BHR/TNP, PCK/KRO, OGS/MIY, RADA/VYMI, TGI/RLY, VPU/APD, TDW/FOXF, SHEN/AP, NVTA/DUK, CTT/BKCC, CERN/PCY, GBAB/FPA, BOOM/MHO, FTEK/RMD, EGLE/RBC, MHN/SSBI, JPC/C, LCI/XAR, CDMO/OTLK, JPM/VCSH, TMST/PJH, BCV/XME, LLNW/BHR, VBLT/SXC, ELTK/SHLO, HIX/ENZL, ACY/PJT, LDOS/EMD, PWB/DUG, KPTI/SAIC, PSHG/VNO, PTSI/FPA, UBSI/NOK, APOG/WAB, DXJ/SMLP, SBBP/DCOM, MC/DLR, GNTX/IPAR, TGT/NTWK, FAZ/KAI, PFO/CHSCL

Control set pairs with a cointegrating relationship:

CINR/BWZ, BIMI/AGTC, ENOR/ATHX, SEED/AVK, FUL/CORR, UG/ITRI, WSC/PED, HTHT/MRIN, KPTI/SAIC

Directors set pairs:

NC/HY, UA/UAA, GOOGL/GOOG, PMT/PFSI, UBP/UBA, SBBX/S, OLP/BRT, BGT/HYT, CTI/CTIC, BGR/BTZ, BGR/BGT, BGR/BGY, BGR/HYT, BGR/BDJ, BGR/BME, BGR/BLW, BTZ/BGT, BTZ/BGY, BTZ/HYT, BTZ/BDJ, BTZ/BME, BTZ/BLW, BGT/BGY, BGT/BDJ, BGT/BME, BGT/BLW, BGY/HYT, BGY/BDJ, BGY/BME, BGY/BLW, HYT/BDJ, HYT/BME, HYT/BLW, BDJ/BME, BDJ/BLW, BME/BLW, ADX/PEO, MPC/MPLX, UMH/MNR, CCL/CUK, MGP/MGM, GF/CEE, RES/MPX, TMQ/NG, ANAT/NWLI, PAA/PAGP, D/DM, EEQ/EEP, UTF/FOF, UTF/RQI, UTF/RNP, FOF/RQI, FOF/RNP, RQI/RFI, RQI/RNP, DO/L, CAT/BA, ABBV/ABT, ENLC/ENLK, GMZ/GER, VBIV/OPK, DISCA/DISCK, TTI/CCLP, KEYS/A, MFM/MIN, MFM/CMU, MFM/MGF, MFM/MMT, MFM/MFV, MFM/CXE, MFM/MCR, MFM/CXH, MFM/CIF, MIN/CMU, MIN/MGF, MIN/MMT, MIN/MFV, MIN/CXE, MIN/MCR, MIN/CXH, MIN/CIF, CMU/MGF, CMU/MMT, CMU/MFV, CMU/CXE, CMU/MCR, CMU/CXH, CMU/CIF, MGF/MMT, MGF/MFV, MGF/CXE, MGF/MCR, MGF/CXH, MGF/CIF, MMT/MFV, MMT/CXE, MMT/MCR, MMT/CXH, MMT/CIF, MFV/CXE, MFV/MCR, MFV/CXH, MFV/CIF, CXE/MCR, CXE/CXH, CXE/CIF, MCR/CXH, MCR/CIF, CXH/CIF, DXC/HPE, GPC/SO, PNC/CSX, SHLM/OMN, FNF/FIS, MKC/TROW, RSG/ECL, RSG/AN, TTNP/CRIS, HYT/DSU, EAT/AZO, PSA/PSB, FCB/CLI, INGR/AKS, PATK/IESC, DE/MMM, PSA/AMH, MUR/MUSA, PULM/XLRN, IBM/PG, VG/SFE, OMN/CMP, MTH/WAL, PMO/PMM, PMO/PIM, PMO/PCF, PMO/PPT, PMM/PIM, PMM/PCF, PMM/PPT, PIM/PCF, PIM/PPT, PCF/PPT, EPC/BLL, TILE/KNL, CTXS/LOGM, SRG/LE, UA/IRDM, UAA/IRDM, UTF/LDP, UTF/RFI

Directors set pairs with a cointegrating relationship:

GOOGL/GOOG, OLP/BRT, HYT/BLW, RQI/RFI, DISCA/DISCK, MFM/CXE, MIN/MGF, MMT/MCR, MFV/CIF, CXE/CXH, PIM/PCF

Employees set pairs:

NC/HY, UA/UAA, GOOGL/GOOG, SBBX/S, UBP/UBA, PFSI/PMT, ADX/PEO, DISCA/DISCK, BRT/OLP, ENLK/ENLC, CTI/CTIC, PCF/PMO, BGR/BDJ, BGR/BGY, BGY/BDJ, PCI/PKO, PCI/PDI, PCI/PCM, PDI/PKO, PDI/PCM, BGT/HYT, PMM/PCF, PMM/PIM, PMM/PPT, PMM/PMO, PIM/PCF, PIM/PPT, PIM/PMO, PCF/PPT, BTZ/HYT, MPLX/MPC, PKO/PCM, PPT/PMO, BME/BGY, PCK/PMF, PCK/PDI, PCK/PNF, PCK/PGP, PCK/PKO, PCK/PNI, PCK/PCN, PCK/PML, PCK/PFN, PCK/PTY, PCK/PMX, PCK/PZC, PCK/PYN, PCK/PCQ, PCK/PHK, PCK/PCI, PCK/PFL, PCK/PCM, PNI/PNF, PNI/PGP, PNI/PKO, PNI/PCN, PNI/PML, PNI/PFN, PNI/PTY, PNI/PMX, PNI/PZC, PNI/PYN, PNI/PCQ, PNI/PHK, PNI/PCI, PNI/PFL, PNI/PCM, PNI/PMF, PNI/PDI, PNF/PDI, PNF/PGP, PNF/PKO, PNF/PCN, PNF/PML, PNF/PMX, PNF/PFN, PNF/PTY, PNF/PZC, PNF/PYN, PNF/PCQ, PNF/PCI, PNF/PHK, PNF/PFL, PNF/PCM, PNF/PMF, ASG/USA, PCI/RCS, BLW/BGY, BLW/BGR, BLW/BTZ, BLW/BGT, BLW/BME, BLW/HYT, BLW/BDJ, PGP/PYN, PGP/PCQ, PGP/PCI, PGP/PHK, PGP/PFL, PGP/PCM, PGP/PMF, PGP/PDI, PGP/PKO, PGP/PML, PGP/PCN, PGP/PMX, PGP/PFN, PGP/PTY, PGP/PZC, PDI/PMF, PDI/PCN, PDI/PML, PDI/PFN, PDI/PTY, PDI/PMX, PDI/PZC, PDI/PYN, PDI/PCQ, PDI/PHK, PDI/PFL, PDI/RCS, BGT/BME, BGT/BDJ, BGT/BGY, BGT/BGR, BGT/BTZ, PML/PKO, PML/PCN, PML/PMX, PML/PFN, PML/PTY, PML/PZC, PML/PYN, PML/PCQ, PML/PCI, PML/PHK, PML/PFL, PML/PCM, PML/PMF, PYN/PCQ, PYN/PCI, PYN/PHK, PYN/PFL, PYN/PCM, PYN/PMF, PYN/PKO, PYN/PCN, PYN/PMX, PYN/PFN, PYN/PTY

Employees set pairs with a cointegrating relationship:

GOOGL/GOOG, DISCA/DISCK, ENLK/ENLC, PDI/PKO, PDI/PCM, PMM/PMO, PIM/PCF, PKO/PCM, PCK/PGP, PCK/PML, PCK/PZC, PNI/PHK, BLW/HYT, PGP/PCQ, PDI/PCN, PDI/PTY, PDI/PFL, PYN/PHK

B Method 2: number of cointegrated pairs in each sample

Control set (random): [0, 4, 1, 2, 1, 1, 2, 1, 2, 0, 1, 0, 1, 2, 1, 2, 1, 1, 1, 2, 2, 3, 2, 1, 2, 1, 0, 1, 6, 1]

Directors set (at least 1 shared attribute): [1, 3, 1, 1, 1, 1, 3, 1, 0, 1]
Directors set (at least 3 shared attributes): [2, 1, 0, 1, 0, 1, 0, 1, 0, 0]
Directors set (at least 5 shared attributes): [2, 1, 2, 1, 0, 1, 0, 1, 0, 0]

Employees set (at least 1 shared attribute): [0, 0, 0, 1, 0, 0, 1, 0, 0, 1]
Employees set (at least 3 shared attributes): [1, 1, 0, 2, 3, 2, 4, 1, 2, 0]
Employees set (at least 5 shared attributes): [1, 2, 2, 2, 2, 3, 2, 2, 1, 3]

Bibliography

[1] M. P. Murray, ‘A drunk and her dog: An illustration of cointegration and error correction’, 1, vol. 48, Taylor & Francis, 1994, pp. 37–39.

[2] N. Shadbolt, T. Berners-Lee and W. Hall, ‘The semantic web revisited’, *IEEE intelligent systems*, vol. 21, no. 3, pp. 96–101, 2006.

[3] T. Berners-Lee, J. Hendler and O. Lassila, ‘The semantic web’, *Scientific american*, vol. 284, no. 5, pp. 34–43, 2001.

[4] M. Uschold and M. Gruninger, ‘Ontologies: Principles, methods and applications’, *The Knowledge Engineering Review*, vol. 11, no. 2, pp. 93–136, 1996.

[5] A. U. Frank, ‘Tiers of ontology and consistency constraints in geographical information systems’, *International Journal of Geographical Information Science*, vol. 15, no. 7, pp. 667–678, 2001.

[6] B. Smith, W. Ceusters, B. Klagges, J. Köhler, A. Kumar, J. Lomax, C. Mungall, F. Neuhaus, A. L. Rector and C. Rosse, ‘Relations in biomedical ontologies’, *Genome biology*, vol. 6, no. 5, R46, 2005.

[7] R. J. Elliott, J. V. D. Hoek and W. P. Malcolm, ‘Pairs trading’, *Quantitative Finance*, vol. 5, no. 3, pp. 271–276, 2005.

[8] E. Gatev, W. N. Goetzmann and K. G. Rouwenhorst, ‘Pairs trading: Performance of a relative-value arbitrage rule’, *The Review of Financial Studies*, vol. 19, no. 3, pp. 797–827, 2006.

[9] N. Huck and K. Afawubo, ‘Pairs trading and selection methods: Is cointegration superior?’, *Applied Economics*, vol. 47, no. 6, pp. 599–613, 2015.

[10] C. W. J. Granger, ‘Developments in the study of cointegrated economic variables’, *Oxford Bulletin of Economics and Statistics*, 1986.

[11] R. F. Engle and C. W. J. Granger, ‘Co-integration and error correction: Representation, estimation, and testing’, *Econometrica*, vol. 55, no. 2, pp. 251–276, 1987.

[12] D. A. Dickey and W. A. Fuller, ‘Distribution of the estimators for autoregressive time series with a unit root’, *Journal of the American Statistical Association*, vol. 74, no. 366, pp. 427–431, 1979.

[13] S. Johansen, ‘Estimation and hypothesis testing of cointegration vectors in gaussian vector autoregressive models’, *Econometrica*, vol. 59, no. 6, pp. 1551–1580, 1991.

[14] B. Do, R. Faff and K. Hamza, ‘A new approach to modeling and estimation for pairs trading’, in *Proceedings of 2006 financial management association European conference*, Citeseer, 2006, pp. 87–99.

[15] J. Gonzalo and T.-H. Lee, ‘Pitfalls in testing for long run relationships’, *Journal of Econometrics*, vol. 86, no. 1, pp. 129–154, 1998.

[16] T. A. Khan, ‘Cointegration of international stock markets: An investigation of diversification opportunities’, *Undergraduate Economic Review*, vol. 8, no. 1, p. 7, 2011.

[17] M. Raddant and D. Y. Kenett, ‘Interconnectedness in the global financial market’, *OFR WP*, pp. 16–09, 2016.

[18] S. R. Baker, N. Bloom, S. J. Davis, K. Kost, M. Sammon and T. Viratyosin, ‘The unprecedented stock market reaction to covid-19’, 2020.

[19] O. Issing, *The globalisation of financial markets*, Remarks by Professor Otmar Issing on 12 September 2000., 2000. [Online]. Available: https://www.ecb.europa.eu/press/key/date/2000/html/sp000912_2.en.html.

[20] E. F. Fama, ‘Efficient capital markets: A review of theory and empirical work’, *The Journal of Finance*, vol. 25, no. 2, pp. 383–417, 1970, ISSN: 00221082, 15406261. [Online]. Available: http://www.jstor.org/stable/2325486.

[21] *Yfinance*, https://github.com/ranaroussi/yfinance, Accessed: 2020/05/09.

[22] *Statsmodels.tsa.stattools.coint*, https://www.statsmodels.org/dev/generated/statsmodels.tsa.stattools.coint.html, Accessed: 2020/05/09.

[23] *Statsmodels.tsa.vector_ar.vecm.coint_johansen*, https://www.statsmodels.org/dev/generated/statsmodels.tsa.vector_ar.vecm.coint_johansen.html, Accessed: 2020/05/09.

[24] J. G. MacKinnon, ‘Approximate asymptotic distribution functions for unit-root and cointegration tests’, *Journal of Business & Economic Statistics*, vol. 12, no. 2, pp. 167–176, 1994, ISSN: 07350015. [Online]. Available: http://www.jstor.org/stable/1391481.

[25] *Nasdaq traded*, ftp://ftp.nasdaqtrader.com/symboldirectory/nasdaqtraded.txt, Accessed: 2020/05/09.

[26] Y. Liu, Q. Zeng, H. Yang and A. Carrio, ‘Stock price movement prediction from financial news with deep learning and knowledge graph embedding’, in *Pacific Rim Knowledge Acquisition Workshop*, Springer, 2018, pp. 102–113.