Department of Computer Science

Submitted in part fulfilment for the degree of BSc.

Exploring Potentially Relevant Relations Between Companies Within Web Ontologies for Stock Trading

Dominic Taylor

29th April 2019

Supervisor: Dimitar Kazakov

Contents

Abstract

In this study we test whether there exists a link between: the presence of certain relations within web ontologies about pairs of companies and whether this the pair's stock price share a significant cointegration relationship. In pairs trading cointegration is used as a proxy measure to gauge how close to one another a pair of companies are. Leading us to develop methods to identify if there exists a significant relationship between companies that share this cointegration relationship and also share relationships in web ontologies.

The study uses minute interval data collected from the NASDAQ exchange for 48 stock symbols. Each of these symbol's also have their company's DBpedia page identified which is then used to identify relations between the companies. Then 2 methods are developed for attempting to predict a relationship between information stored in the web ontology and the cointegration of a pair of stocks. Method 1 considers the volume of distance 1 and distance 2 relations between the companies and the second method tries to identify important relations that their presence alone suggests that the companies may be cointegrated.

From this study we have been unable to identify a significant link between the information stored about a pair of companies in web ontologies and the companies' cointegration. This in itself does not mean the methods themselves are flawed. Potentially by expanding the number of stock symbols used using a more data rich ontology it may be possible to retry these methods and find a link.

Executive Summary

This report investigates whether or not pairs of companies traded on the stock market which share various institutional and personal links are more likely to have their stock prices share a cointegration relationship than those that do not share these links. The report identifies 48 companies traded on the NASDAQ exchange which also have a DBpedia page. DBpedia is a web ontology which contains facts about entities. Using the facts stored in the ontology we attempt to identify relationships between pairs of companies' and then assess if there is a link between companies which have these relations and whether or not the companies stock prices are cointegrated.

These relationships can be seen in the real world thus providing the motivation to further investigate these links. In 2015 during the Volkswagen diesel emissions testing scandal the company's stock rapidly dropped in value. Similar behaviour was seen in the stock prices of Volkswagen's subsidiaries: Audi and Porche. By recognising these important relationships one may be better able to make investment decisions as you are more aware of how changes in one stock value may lead to changes in another. Additionally, Qualcomm (QCOM) and Intel (INTC), both large semi-conductor companies, were found to have a significant cointegration relationship: raising the question that does being a member of the same industry have an affect on the likelihood of a pair of stocks being cointegrated?

Cointegration is an important measure for selecting pairs of stocks to use in the Pairs Trading investment strategy, it is used as a proxy measure for gauging the closeness of two companies. Therefore we attempt to identify if there is a link between companies' cointegration and relations found in web ontologies which if successful could potentially form part of a new strategy for selecting companies to perform pairs trading with. This study does not go as far as to investigating how this new technique could be applied to pairs trading.

We hypothesise that if a pair of stocks are highly cointegrated then they shall have some relations and properties which can be identified using web ontologies. Thus, there shall be a statistically significant difference in the cointegration of pairs of stocks with these relations as opposed to those which do not share them.

To test our hypothesis minute interval stock market data was collected

from the Alpha Vantage API [1] for the 48 chosen stock symbols over the period 2019/04/12 09:31:00 - 2019/04/18 16:00:00. Cointegration 'pvalues' were then calculated for each pairwise combination of the stocks and two methods were developed for calculating a coefficient for predicting if a pair of companies will be cointegrated.

Most of the companies used in the study were identified automatically by finding DBpedia pages of companies that had the relation 'http://dbpedia.org /property/symbol' to a symbol known to be traded on the NASDAQ market. Whilst the approach did work for some companies many of the companies that are traded on the NASDAQ exchange did not have a DBpedia page and a smaller subset of those had the required relation. Therefore some companies DBpedia pages were identified manually.

The first method is based on the concept that highly cointegrated pairs may share many relationships within the web ontology compared to pairs without a cointegration relationship. Therefore by counting the number of distance 1 and 2 relationships between the companies one can predict whether or not a pair of companies are likely to be highly cointegrated.

The second method attempts to identify relationships between pairs of companies where the presence alone of that relation suggests the pair may have a significant cointegration relationship. This was done by splitting the data into two sets: an Identification set and Test set. The Identification set was used to select any relationships between pairs of companies and then the mean cointegration of pairs of companies that hold that relation was calculated. Subsequently the relations with a significant mean 'pvalue' were selected and used to construct queries to identify if pairs of companies in the Test set which held these relations had a statistically significantly higher mean cointegration than those that did not have these relations.

Having performed the investigation neither method provided significant results. However that is not to say that the methods are ineffective in themselves. The web ontology often did not contain key information about companies, highlighted by the fact that some companies' pages did not even contain their stock symbol. Therefore relationships we know to exist in the real world could not be identified through the ontologies, leading to it being difficult to calculate a good cointegration predictor or identify significant relations within the ontology. Had we access to more information rich ontologies, potentially, the methods trialled could prove more successful. Therefore, we suggest that in further work these methods are repeated but using a larger set of stock symbols and more data rich ontologies.

There were no legal, social, ethical, professional, or commercial issues identified as part of this study.

List of Figures

List of Tables

1 Introduction

1.1 Introduction

This project brings together concepts from multiple fields including Computer Science and Economics. We want to test the hypothesis that information stored in web ontologies can help to identify highly cointegrated pairs of stocks. If this proves to be true, this information could help in selecting pairs of stocks which would make good candidates for pairs trading.

We believe that the open data stored in web ontologies, such as DBpedia, could assist us in being able to identify pairs of stocks which are highly cointegrated by finding common properties, such as institutional or personal links, shared by pairs of stocks. For instance we believe that if one company is a subsidiary of another then they are more likely to have highly cointegrated stock values as their finances depend on one another and therefore we can use the property 'http://dbpedia.org/ontology/subsidiary' from DBpedia to identify pairs of companies that are parent company and subsidiary. Then, using historical stock market data we may test the hypothesis.

Pairs trading is a form of Statistical Arbitrage which relies on exploiting a divergence in the price of a pair of stocks from a common equilibrium. Cointegrated stocks have a historic trend of moving together and when they do diverge they should return to one another, eventually. It is during these divergences that one may try to make a profit. When choosing companies to perform this trading strategy with cointegration is often used to measure a pair's suitability. Cointegration can be used as a proxy measure for gauging how close to each other a pair of companies are. As the technique has become more widely used it has become less effective. If these methods prove successful, further work, could use the information identified as important in web ontologies to help inform decisions made when using this strategy, potentially augmenting the effectiveness of Pairs Trading.

On the NASDAQ exchange alone, at the time of writing, there are 3414 [2] stocks being traded and therefore to find a cointegration coefficient for every single pair would require $\binom{3414}{2} = 5825991$ tests to be carried out. Therefore if we can find a method to effectively find pairs that are likely to be cointegrated we can significantly reduce the number of tests needed to be carried out.

Considering this, we hypothesise that there exists some link between relationships stored within web ontologies, about pairs of companies, and whether or not these companies share a significant cointegration method. Therefore, we shall develop a method for identifying these relationships and test to see if there is a statistically significant difference in the level of cointegration of pairs that share these relations compared to those that do not. This project shall be considered a success if we are able to identify if such a link exists.

2 Literature Review

This chapter will review the three concepts central to this piece of work: Cointegration as a way of establishing a link between the prices of two stocks; Pairs Trading as a strategy that monetises the information provided by cointegration and web ontologies as an additional source of information that may improve the performance of a Pairs Trading strategy.

2.1 Cointegration

Two time series datasets are cointegrated if there exists some linear combination of them which results in a stationary time series. In the paper 'Developments in the study of cointegrated economic variables' [3] Granger states for two time series, x_t and y_t, to be cointegrated, given constant A, the linear combination of them:

$$z_t = x_t - Ay_t$$

Must result in a time series z_t which is stationary, i.e. integrated of order 0 ($I(0)$). This resulting series will have a mean which the series will rarely diverge from and it will frequently cross this mean value. To test for cointegration first we must find the order of integration of the two time series x_t and y_t using the Dickey-Fuller test and provided both x_t and y_t are of order $I(1)$, i.e. the null hypothesis H_0 is accepted, then proceed to compute the cointegration regression:

$$x_t = c + \alpha y_t + a_t$$

Where a_t is the linear combination of x_t and y_t and if they are cointegrated then a_t will have an order of integration of $I(0)$. Using this method has shown that in the United State's National Income is cointegrated with Consumption [4] which therefore suggests that there is a relationship between income and expenditure.

2.2 Pairs Trading

Pairs Trading is a form of Statistical Arbitrage, a set of trading strategies which involve buying and selling stocks simultaneously exploiting the price

difference in order to make a profit [5]. Gatev et al. [6] describe Pair Trading as simple strategy that relies on finding a pair of stocks that have historically moved together and when they have historically diverged eventually they return to one another. Therefore, in the future when the values of the stocks diverge short, (sell), the higher priced stock and buy the lower one. Eventually the prices, of the two stocks, should converge and by selling the previously under-priced stock one can make a profit.

Huck et al. [7] discusses how cointegration can play a key roll in selecting pairs of stocks for pairs trading. If a pair of stocks are cointegrated then they have shared a long-term equilibrium meaning that in the past if their prices have diverged then eventually they have returned and as discussed earlier this property is critical for successfully executing the Pairs Trading strategy. Furthermore, Huck's research suggested that, compared to other methods, there was significantly higher chance of pairs selected using cointegration eventually converging.

Like Huck, in McSharry's report [8] he outlines a strategy for effectively selecting pairs for trading also using cointegration. The report suggests that using correlation would not be a suitable measure for finding pairs to trade as high-correlation will lead to little divergence between the stocks and therefore few opportunities to exploit to make a profit. Instead McSharry proposes to use cointegration to measure pairs' suitability as it allows for the values to diverge from a mean which can then then be exploited for profit.

2.3 Web Ontologies

The semantic web is a technology for allowing machines to understand the meaning behind content on the World Wide Web therefore enabling them to perform intelligent tasks, such as booking appointments, automatically [9]. This is in opposition to how the World Wide Web works currently: by linking together documents but the machines having no comprehension of, and therefore cannot utilise, the contents within the documents. Burners-Lee et al. go on to describe how Ontologies, a collection of statements that relate entities to subjects by some predicate (e.g. entity: "10 Downing Street"; Predicate; "has Postcode"; Subject: "SW1A 2AB"), can be used by machines to find relationships between entities and infer information.

These ontologies may be created and hosted online by anyone: the Linked Open Data Cloud [10] project maintains a collection of ontologies that are all interlinked. Using these interlinked ontologies one may find information about one entity spanning many ontologies and can use the relations between ontologies to link these facts.

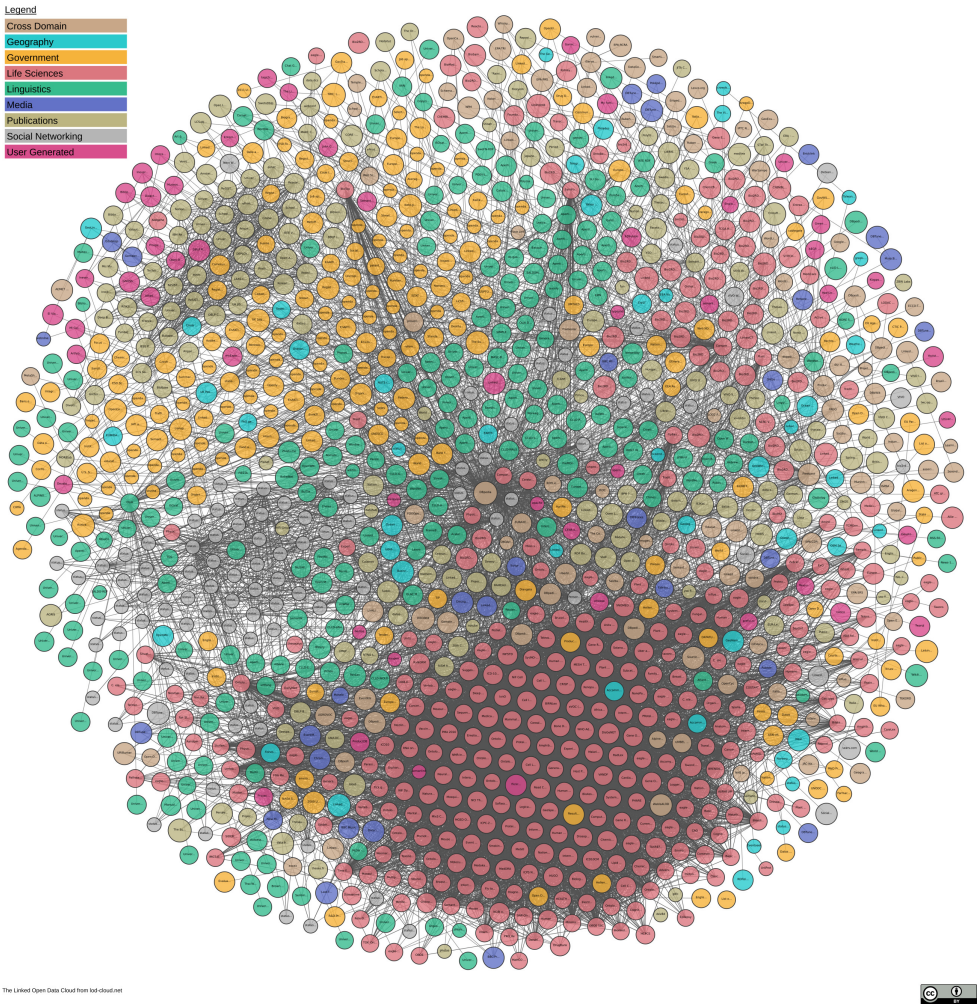

Figure 2.1: Network map of the Linked Open Data Cloud [10]

Qu et al. [11] takes a novel approach to using ontologies: by using them to store real time data which can be annotated with additional information using extra relations within the ontology. They have constructed an ontology that contains events which are related to the value of a stock at that time stamp and tweets made also at that timestamp. The research focused on the time period in and around the Volkswagen diesel emissions testing scandal whereby they attempted to find links between twitter data and the stock price by using SPAQRL queries. Using these queries the researchers could extract tweets that coincided with a drop, of more than 1%, in the stock's value and using an additional query they could find the keywords used when the stock price dropped by more than 2%. Finding these links using more traditional data representations would likely be more complicated as it would require working with multiple datasets, i.e. one of stock prices and one of tweets, and integrating them together. Finally, a possible direction for future work is proposed whereby public Ontologies, e.g. DBpedia, are used to automatically link one company to another

and then assess whether these related companies also feel the effects of adverse events experienced by the company they are linked to.

Whilst Qu et al.'s approach to integrating semantic ontology data with time series data is unusual it is not the only instance of attempting this. Božić et al. [12] have taken the approach of integrating time series data with meta-data stored in an ontology in order to be able to make connections between the meta-data of the ontology and the time series data. In this instance a Time Series Processor uses the ontology to inform how to attach relevant meta-data to the time series data which is similar to how Qu et al. linked the time series stock data to tweets made at that time. However, where they differ is that Qu et al. store the data and tweets together in an ontology whereas Božić et al.'s outputs another time series annotated with the semantic data.

Knowledge graphs can be used to model the relations between entities in the financial and economic domain. For instance, they can graph the relationships between entities such as: companies, management and news events and this information can be an effective tool for investors to use to inform their decisions. Liu et al. [13] discus how whilst there have been some application of knowledge graphs to model events in financial news many of the approaches have failed to capture the semantic meaning of the events and therefore it is difficult to automatically infer the effect of an event on a company's stock price, instead they just store the event as a 'bag of words'. Liu et al. propose using convolutional neural networks in order to extract the semantic meaning of financial news headlines and then store this meaning as parts of tuples in a knowledge graph. For instance, from the headline 'Apple stops selling some devices online in Germany' one could extract the tuple ('Apple', 'stops selling', 'some devices') and by breaking the headline down into a tuple containing a subject, predicate and object we could now build a system for identifying the semantic meaning of the headline. The focus of the paper is the implementation of a deep learning method to extract this semantic information, over a traditional machine learning method and, the results show that using this method is effective and can produce better results than traditional machine learning.

3 Motivation and Problem Analysis

3.1 Motivation

This project relies on being able to establish relations between companies with the prospect that these relations indicate the presence of high cointegration between two companies stock values. To demonstrate the potential of this system examples of companies that demonstrate that these links and their stock prices show a relationship are cited.

3.1.1 Volkswagen, Audi and Porche: Emissions testing scandal

In 2015 Volkswagen were found to be using a 'defeat device' in some of the diesel cars they produce in order to fool laboratory testing into believing their cars released fewer emissions than they actually did [14]. In reality some of Volkswagen's cars were producing far more emissions than regulations allow leading to them being heavily fined. The revelation of this news lead to drops in Volkswagen's share price [15].

As the value of Volkswagen's stock dropped so did the stock values of Porche and Audi. Audi and Porche are both part of the Volkswagen Group [16] and one could ask the question of whether this relationship contributed to the stock prices sharing similar movements. By being aware of this information one may be able to make better informed investment decisions.

3.1.2 Intel and Qualcomm

Using minute interval, high, stock prices from the period 2019/04/12 09:31:00 - 2019/04/18 16:00:00 of Intel (INTC) and Qualcomm (QCOM), shown in figure 3.1, a cointegration 'pvalue' of 0.04227 was calculated suggesting that a significant cointegration relationship exists. Both Intel and Qualcomm are members of the semi-conductor industry and are competing for the same market. Having these relationships and being a member of the same

industry may be a factor in why these companies share a cointegration relationship.

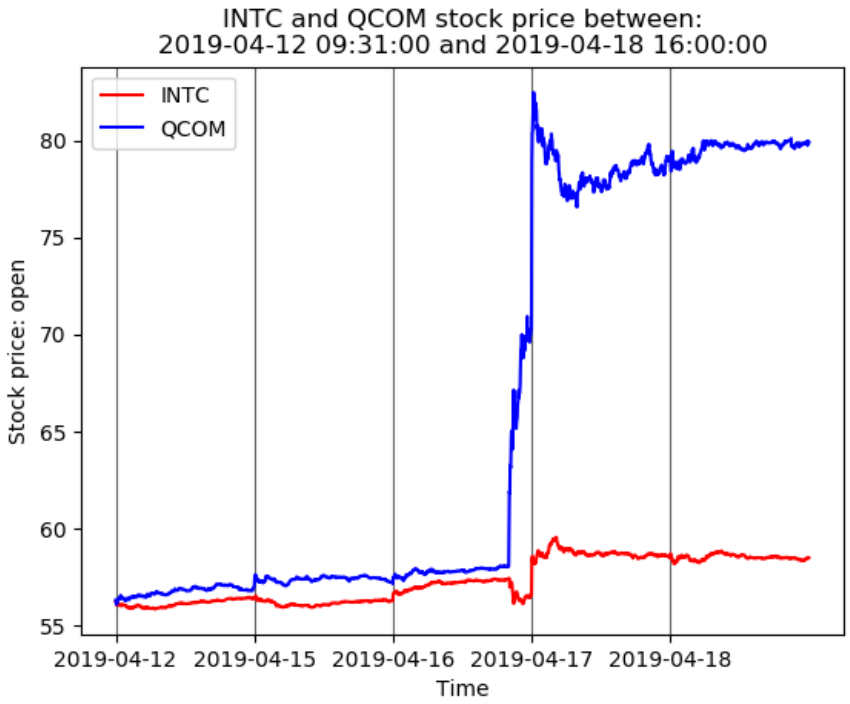

Figure 3.1: Minute interval, open, stock price of Intel (INTC) and Qualcomm (QCOM) on the NASDAQ exchange for the period 2019/04/12 09:31:00 - 2019/04/18 16:00:00

3.1.3 Alphabet

On the NASDAQ exchange Alphabet, the parent company of Google, is traded under more than one symbol. Its class A stock under 'GOOGL' and class C under 'GOOG'. Testing for cointegration in the period 2019/04/12 09:31:00 - 2019/04/18 16:00:00 resulted in a 'pvalue' of 0.02343 indicating a significant cointegration relationship. A possible explanation as to why the two stocks' prices share a relationship may be as they are from the same company. If we could identify if two stock symbols came from the same company then this could be an important factor in predicting if the companies are likely to have a cointegration relationship.

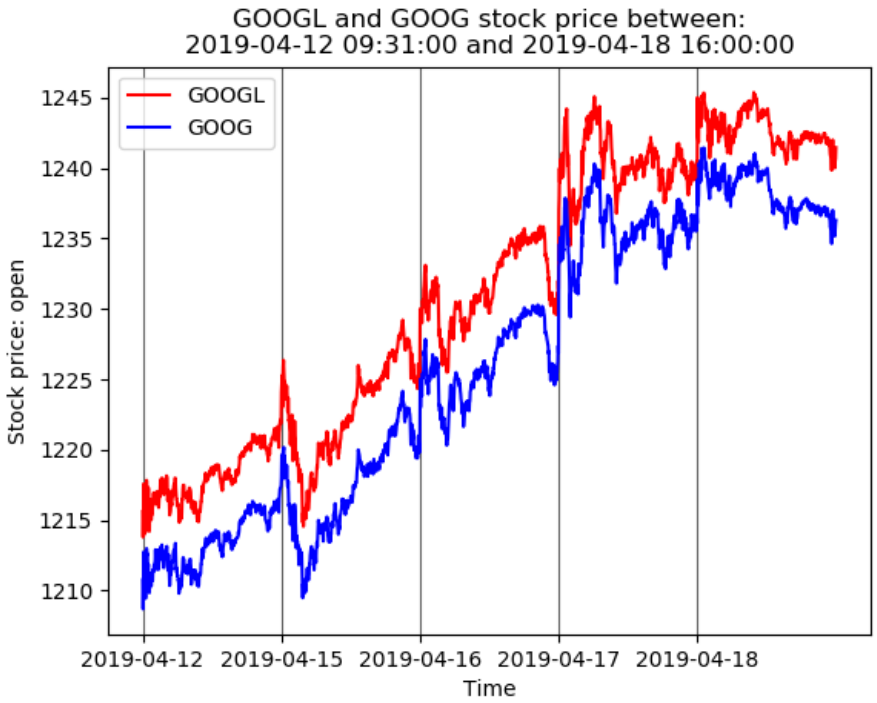

Figure 3.2: Minute interval, open, stock price of Alphabets class A (GOOGL) and class C (GOOG) stock prices on the NASDAQ exchange for the period 2019/04/12 09:31:00 - 2019/04/18 16:00:00

3.2 Problem Analysis

Having considered the motivation this leads to the concept that: pairs of companies where various personal and institutional relationships exist there may also exist a relationship between the pair's stock prices. We suggest that for pairs with these links a significant cointegration relationship could exist and we believe the information we need to find these links is stored within public Web Ontologies.

The overlapping area of Figure 3.3 contains pairs of companies that have a significant cointegration relationship and also share some links stored in Web Ontologies. Our investigation intends to see if we are able to develop a method for automatically identifying these companies which lie in that area.

Pairs trading uses cointegration as a proxy measure for gauging how closely related a pair of companies are. We shall also use cointegration to measure how close a pair of companies are and intend to develop a method which calculates the closeness of a pair of companies using information within web ontologies.

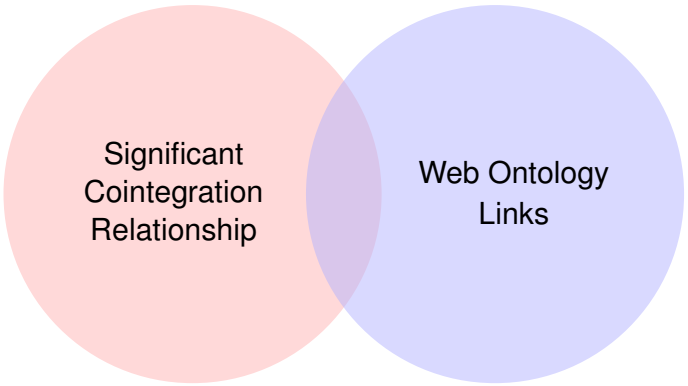

Figure 3.3: Venn diagram showing the possible properties a pair of stocks may hold

Whilst we do not intend to go as far as looking at how our newly developed measure could be used to identify companies for pairs trading; potentially if our method proves successful one could then go on to looking at how well this new measure works for selecting pairs to use in pairs trading.

3.3 Hypothesis

Knowing that cointegration can be used to measure the closeness of two companies and that their exists relations about companies traded on the stock market within web ontologies: we hypothesise that if a pair of stocks are highly cointegrated then they shall have some relations and properties which can be identified using web ontologies. Thus, there shall be a statistically significant difference in the cointegration of pairs of stocks with these relations as opposed to those which do not share them.

4 Method

4.1 Method Overview

To test the hypothesis a set of companies shall be chosen which are traded on the NASDAQ stock exchange; then using minute interval stock market data a cointegration test shall be performed on pairs of stocks and a pvalue calculated to indicate the likelihood of the pair's stock price being cointegrated.

The stock symbols used in the investigations must all have a DBpedia page so that the web ontology may be queried in order to source additional information about these companies. These DBpedia pages shall be identified automatically, however additional pages maybe added manually if they have failed to be identified.

Using the web ontologies: for each pair of companies a coefficient shall be calculated which indicates how likely we believe the pair to be cointegrated. Two methods shall be used to calculate this coefficient: one looking at the volume of relations between a pair of companies and another based on identifying the presence of significant relations.

After having calculated the cointegration 'pvalue' and the coefficient indicating our belief that the pair are cointegrated: statistical tests shall be used to see if there is a statistically significant difference between companies that have relationships identified in the web ontology and have a cointegration relationship compared to those that do not share these ontology links.

Figure 4.1 shows a diagrammatic overview of the method used and full details of the method are available in section 4.2.

4.1.1 Method Diagram

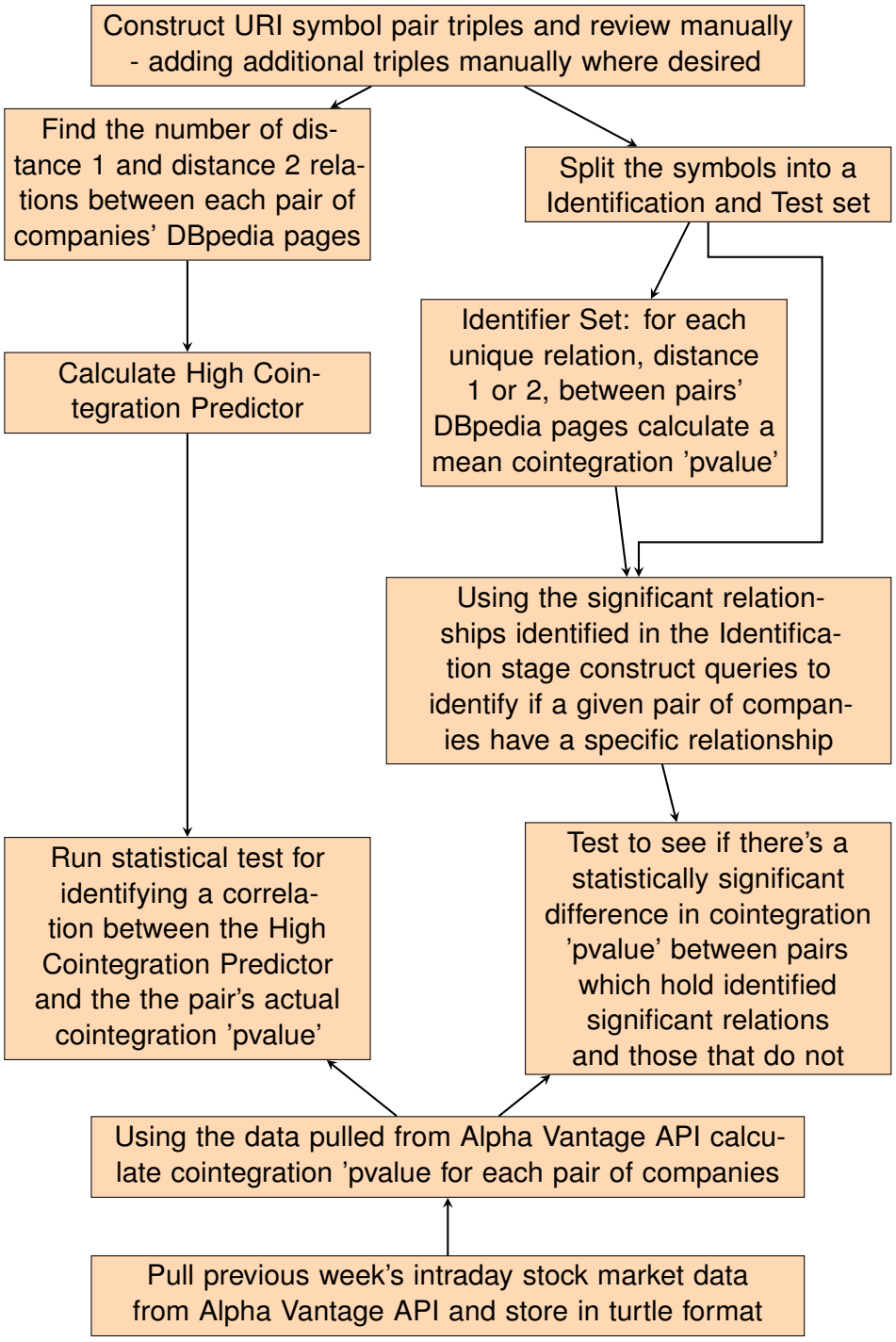

Figure 4.1: Diagram showing an overiew of the method

```
SELECT ?company ?symbol WHERE {
   ?company <http://dbpedia.org/property/symbol>
   "{symbol}"^^rdf:langString .
   ?company <http://dbpedia.org/property/symbol>
   ?symbol .
}
```

Figure 4.2: Query template for identifying the DBpedia page of a company with the supplied '{symbol}'

4.2 Full Method

4.2.1 Identification of Companies' DBpedia Pages

A list of all the stock symbols of companies traded on the NASDAQ exchange is available from [17]. These symbols are extracted and SPARQL queries are automatically constructed from a template, Figure 4.2, by replacing 'symbol' with an actual stock symbol, e.g. 'MSFT' for Microsoft, and then used to find entities which have the relation 'http://dbpedia.org/property/symbol' to the supplied symbol.

Using the results of these queries a turtle file, 'symbolURIPairs.ttl', is constructed which contains triples mapping from a companies DBpedia URI to their NASDAQ stock symbol. Due to potential inaccuracies in the DBpedia data the turtle file is then reviewed manually to check for incorrect relations and additional relations may be added manually if the automated system has failed to identify them. The manual additions are stored in 'symbolURIPairsManual.ttl'.

4.2.2 Stock Market Data

In order to calculate the cointegration level of pairs of stocks minute interval stock data is used. This data is collected from the Alpha Vantage API, [1], which provides an API for pulling the previous week's minute interval stock price for any given symbol traded on the NASDAQ exchange.

The data is provided in a CSV file with fields: timestamp, high, low, open, close and volume. The CSV file is parsed into a turtle format whereby each row in CSV file is converted into a set of triples, the structure of which is shown in 4.3.

All of these triples are stored in a single file, 'finacialData.ttl', which contains all the data for all symbols. By storing the data in a turtle file SPAQRL queries may easily be constructed which can, for example, show

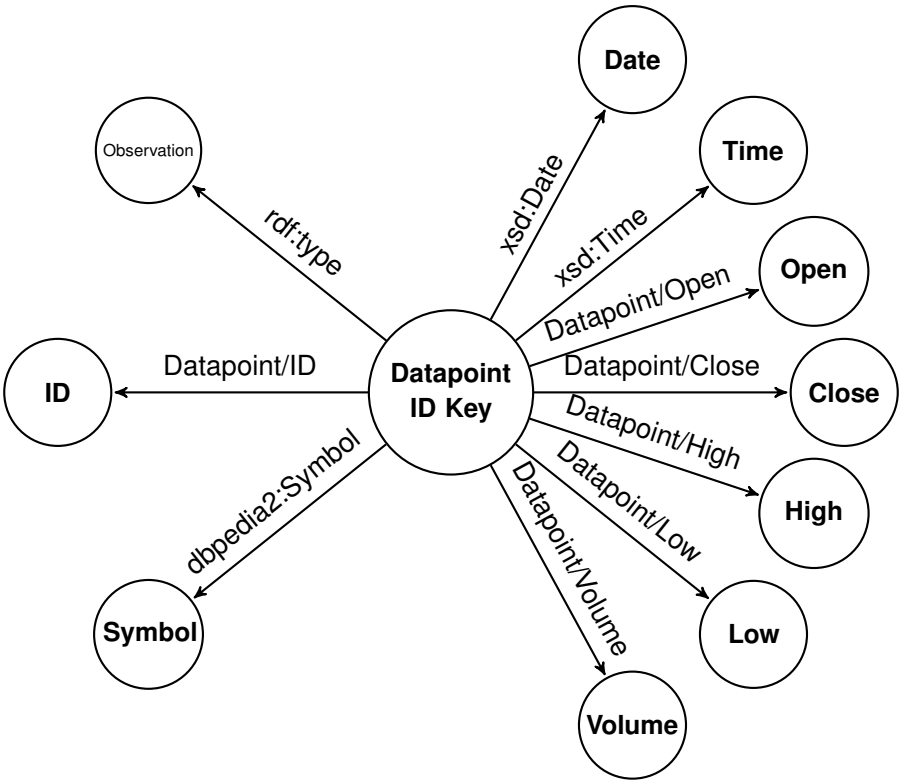

Figure 4.3: Structure of a single datapoint of financial data in RDF format

which stock symbols have data stored about them; return the date range the file covers or extract specific stock data for a collection of symbols.

4.2.3 Calculating Cointegration Values

Using the stock data pulled from the Alpha Vantage API, and converted to turtle format, a cointegration test is run which calculates a 'pvalue' indicating if the two time series are cointegrated. The cointegration test is the Engle-Granger two-step cointegration test which is implemented as part of the Stats Models library for Python [18].

The cointegration test has the null hypothesis that there is no cointegration relationship and an alternative hypothesis that a cointegration relationship exists. Thus a small 'pvalue' means that we may reject the null hypothesis and therefore a cointegration relationship may exist.

A stock price can only be provided when trades occur, which for some popular stocks is every minute, however less popular stocks may not have a price at every minute. In cases of missing data, i.e. one or both stocks do not have a price at a given timestamp, then the entire time stamp shall be discarded for the pair. The test will only be carried out on pairs of stocks

```
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
PREFIX rdf:
<http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX dbpedia2: <http://dbpedia.org/property/>

SELECT DISTINCT ?symbol ?date ?time ?open ?close
?high ?low ?volume WHERE {
   { ?datapoint1 dbpedia2:symbol '{symbol1}}' }
   UNION { ?datapoint1 dbpedia2:symbol '{symbol2}' } .
   ?datapoint1 dbpedia2:symbol ?symbol .
   ?datapoint1 xsd:date ?date .
   ?datapoint1 xsd:time ?time .
   ?datapoint1 <Datapoint/open> ?open .
   ?datapoint1 <Datapoint/close> ?close .
   ?datapoint1 <Datapoint/high> ?high .
   ?datapoint1 <Datapoint/low> ?low .
   ?datapoint1 <Datapoint/volume> ?volume .
} ORDER BY ?time
```

Figure 4.4: SPARQL query used to select the stock prices of two companies, 'symbol1' and 'symbol2' should be replaced with the stock symbols one wishes to extract the data of.

with over 100 shared datapoints.

The cointegration test was performed using the opening price of every minute. This was chosen as it guarantees that the price value is from the exact same time for every datapoint, whereas if using an attribute such as high the value could come from any time within a 1 minute range.

4.2.4 Calculating High Cointegration Likelihood

Method 1: URI Distances

The following method stems from the concept that highly cointegrated stocks are likely to share a higher volume of relations within Web Ontologies than those that are not highly cointegrated. To test this concept two queries shall be constructed: one which counts the number of distance 1 relations between a pair of companies URIs and another which counts the number of distance 2 relations. Distance is defined as the number of steps between the two companies in the Ontology. A distance 1 relation may have one of two forms, as shown in Figure 4.5 and distance 2 relations may take four forms, as shown in Figure 4.6. Longer distance relations were not considered as soon all entities become connected to nearly every other

entity, so these long relations essentially have no unique meaning.

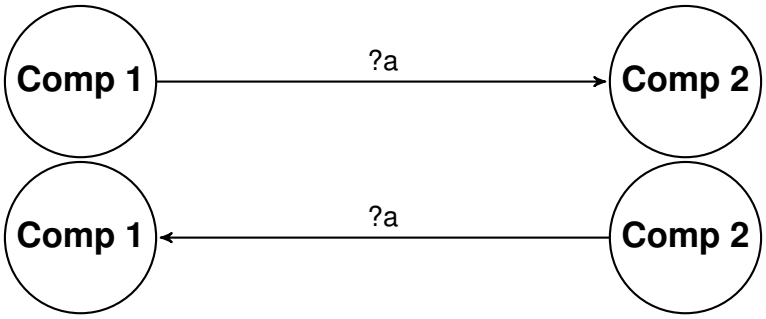

Figure 4.5: All possible distance 1 relations between a pair of company URIs

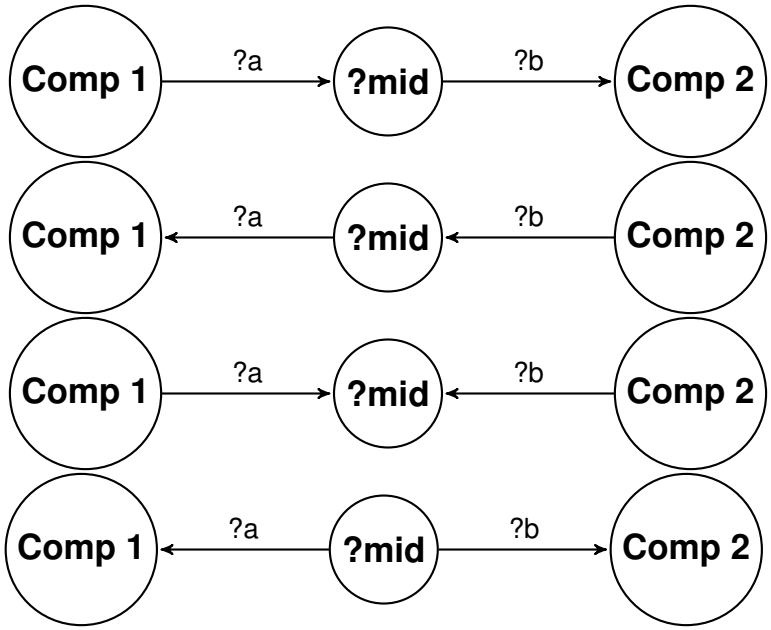

Figure 4.6: All possible distance 2 relations between a pair of company URIs

Each of these queries shall be run for any given pair of companies' DBpedia pages and 2 values shall be returned: the number of distance 1 relations and the number of distance 2 relations. For this method, the type of these relations is not being considered, only the volume of them. These two values shall be combined using:

$$HighCointegrationPredictor = log(DistOneRelations * 100 + DistTwoRelations)$$

$DistOneRelations$ is weighted much heavier than $DistTwoRelations$ in calculating the $HighCointegrationPredictor$ as intuitively shorter distance relations are more significant.

To evaluate the success of the method a correlation test, (the Spearman test), will be performed to identify if there is a signifcant correlation between the $HighCointegrationPredictor$ and the pair's cointegration 'pvalue'. The test will be performed using SciPy's implementation [19].

Method 2: Identifying Significant Relations

Some relationships may be more important in indicating the likelihood of cointegration than others. By identifying what relations are significant we are able to then construct queries which detect the presence of these relations for a specific pair of companies potentially this could suggest whether or not a pair of stocks are likely to be cointegrated.

The data is split into two sets: an Identification set and a Testing set. The Identification set is used to identify which relations are important and the Test set is used to verify if this holds. The Identification set contains 75% of the stock symbols, randomly selected, and the remaining 25% of symbols make up the test set.

The initial identification stage is carried out using a query following the template shown in Figure 4.7, (for finding distance 2 relations - distance 1 is slightly different). Every pairwise combination of symbols in the Identification set shall be tested, the pair's URIs will replace 'companyURI1' and 'companyURI2'.

For every unique relationship found a mean 'pvalue' shall be calculated from the cointegration 'pvalue' of pairs that have this relation. Relations with a mean 'pvalue' of less than 0.05 are deemed significant and are then used to construct SPARQL queries, from a template shown in Figure 4.8, to check if a specific pair of stocks holds a specific relationship. In the template 'companyURI1' and 'companyURI2' will be replaced with the companies' DBpedia URIs and 'relationA' and 'relationC' will be replaced with the relation URIs that we are checking for.

Once the pairs, in the Test set, which hold these significant relations have been identified a statistical test will be run to see if for each of these significant relationships identified in the first stage if there is a statistically significant difference in the mean 'pvalue' of pairs that hold this relationship compared to those that do not.

```
SELECT ?a ?c WHERE {
   {
     <{companyURI1}> ?a ?b .
     ?b ?c <{companyURI2}> .
   } UNION
   {
     <{companyURI2}> ?a ?b .
     ?b ?c <{companyURI1}> .
   } UNION
   {
     ?b ?a <{companyURI2}> .
     ?b ?c <{companyURI1}> .
   } UNION
   {
     <{companyURI2}> ?a ?b .
     <{companyURI1}> ?c ?b .
   }.
   FILTER(?a != rdf:type && ?c !=rdf:type && ?a !=<http
      ://dbpedia.org/ontology/wikiPageWikiLink> && ?c
      !=<http://dbpedia.org/ontology/wikiPageWikiLink>).
}
```

Figure 4.7: SPARQL query template for finding, distance 2, significant relations between pairs of companies DBpedia pages

```
SELECT ?b WHERE {
   {
     <{companyURI1}> <{relationA}> ?b .
     ?b <{relationC}> <{companyURI2}> .
   } UNION
   {
     <{companyURI2}> <{relationA}> ?b .
     ?b <{relationC}> <{companyURI1}> .
   } UNION
   {
     ?b <{relationA}> <{companyURI2}> .
     ?b <{relationC}> <{companyURI1}> .
   } UNION
   {
     <{companyURI2}> <{relationA}> ?b .
     <{companyURI1}> <{relationC}> ?b .
   }.
}
```

Figure 4.8: SPARQL query template for finding if a pair of stocks hold a specific, distance 2, relationship

5 Results

5.1 URI Symbol Pairs

The automated system for identifying pairs of URI and their stock symbols was very ineffective. Using a list of all stock symbols traded on the NASDAQ exchange only 34 DBpedia pages were able to be identified automatically. An additional 15 were added manually.

Furthermore two symbol URI pairs were misidentified: the symbol 'ROSE' was paired with the URI 'http://dbpedia.org/resource/Repression_of_heat_shock_gene_expression_(ROSE)_element' when actually that symbol represents the Rosehill Resources Inc. The symbol 'CFA' was paired with DBpedia page 'http://dbpedia.org/resource/West_African_CFA_franc'. 'CFA' actually represents VictoryShares US 500 Volatility Wtd ETF. Neither of the actual companies have DBpedia pages and therefore they were removed from the list. and therefore this triple was manually removed. Leading to a total of 48 URI symbol pairs being used in this investigation.

The table of all pairs which were added automatically and manually may be found in the appendix at 8.1 and 8.2.

5.2 Data Analysis

The following results were calculated using minute interval data from the NASDAQ exchange collected via the Alpha Vantage API in the period 2019/04/12 09:31:00 - 2019/04/18 16:00:00. Only pairs of stocks with over 100 shared datapoints during the period were considered so as to ensure the calculated cointegration 'pvalue' was reliable, as calculating cointegration with very few datapoints can lead to an illusion of cointegration when in fact it does not exist. The distribution of 'pvalues' is shown in 5.1.

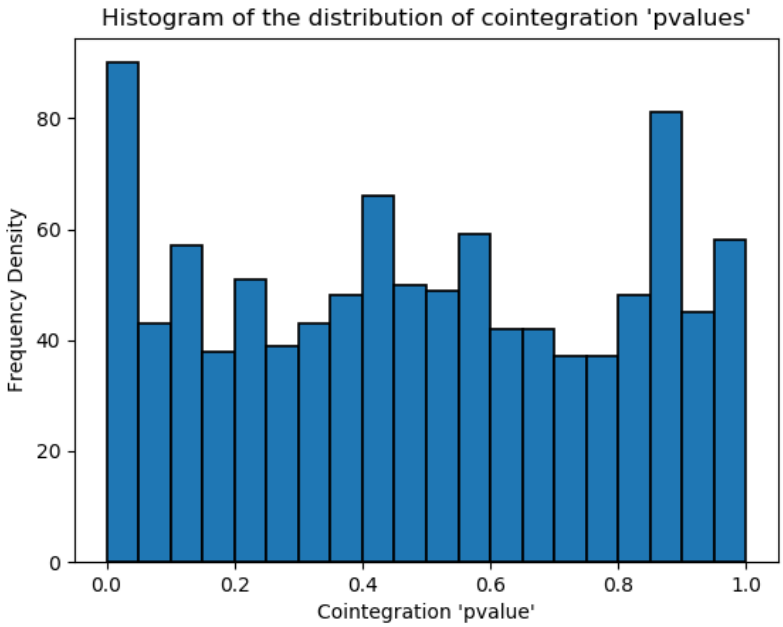

Figure 5.1: Histogram showing the distribution of 'pvalues' for every pairwise combination of stock symbols

5.3 Method 1: URI Distances Results

Having calculated a cointegration pvalue for each pair of stocks and a corresponding 'Ontology distance score' these two values were plotted against one another as shown in 5.2.

From examining the plot it shows no significant trend in the data. At the bottom of the graph shows pairs spanning the full range of possible pvalues, (0 to 1), which all have an Ontology distance score of 0. Furthermore, pairs that did have a score of greater than 0 also span the full range of pvalues.

The results of the Spearman correlation produced a ρ value of 0.1368 and a 'pvalue' of 0.00002. ρ may take a value between -1 and 1 where: a lower value indicates a negative correlation; a higher value a positive correlation and a value close to 0 suggests no correlation. Thus, ρ suggests that there is a very weak positive correlation and the very low 'pvalue' suggests that we may reject the null hypothesis of there being no correlation.

A positive correlation suggests that pairs with higher 'pvalues' also have a higher High Cointegration Predictor. This is the opposite trend to what was expected as a low 'pvalue' indicates the pair are cointegrated. However the trend is so weak that this is likely to just be a quirk of the data and there is in fact no relationship between the pairs cointegration and the calculated High Cointegration Predictor.

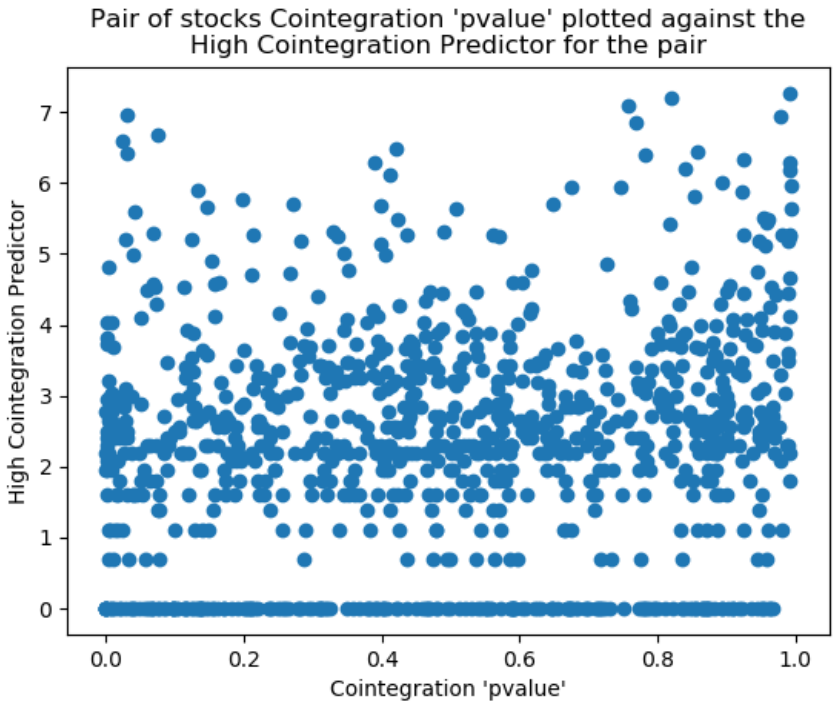

Figure 5.2: A plot showing the pair's 'ontology distance score' against the pair's cointegration pvalue

5.4 Method 2: Identifying Significant Queries Results

The symbols used in the Identification set are listed in the appendix section 8.2.1 and the symbols in the test set are listed in appendix section 8.2.2.

SPARQL queries were used to identify all distance 1 and 2 relations between pairs of companies' DBpedia pages. The cointegration 'pvalues' of pairs of companies which hold any given relation were used to calculate a mean cointegration 'pvalue'. Those relationships with a mean 'pvalue' of lower than 0.1 were considered to be significant.

No significant distance 1 relationships were able to be identified and relatively few distance 2 relationships could be found, those that were found are shown in table 5.1.

By plotting the log frequency of a relation's occurrence against the mean cointegration 'pvalues' of pairs that hold that relation, shown in figure 5.3, one can see that the more relations which hold the mean 'pvalue' tends to 0.5. Suggesting that many of these relations have no significance as if many pairs hold it generally does not have a significant mean.

Table 5.1 shows the significant queries. These were then taken and used

Relation A	Relation B	Mean P Value	Relation Frequency
dbp:incomeYear	dbp:numLocationsYear	0.0137	1
dbp:incomeYear	dbp:numEmployeesYear	0.0137	1
dbp:numEmployeesYear	dbp:numLocationsYear	0.0137	1
dbp:numEmployeesYear	dbp:numEmployeesYear	0.0137	1
dbp:assetsYear	dbp:numLocationsYear	0.0137	1
dbp:assetsYear	dbp:numEmployeesYear	0.0137	1
dbp:netIncomeYear	dbp:numLocationsYear	0.0137	1
dbp:netIncomeYear	dbp:numEmployeesYear	0.0137	1
dbp:numLocationsYear	dbp:revenueYear	0.0137	1
dbp:numEmployeesYear	dbp:revenueYear	0.0137	1
dbp:equityYear	dbp:numLocationsYear	0.0137	1
dbp:equityYear	dbp:numEmployeesYear	0.0137	1
dbp:hqLocationCountry	dbp:hqLocationCountry	0.0137	1
dbo:bioavailability	dbp:width	0.0041	1
dbo:netIncome	dbo:operatingIncome	0.0118	1
dbp:nativeClients	dbp:nativeClients	0.0032	4

Table 5.1: Distance 2 relations with mean 'pvalue' under 0.05. 'http://dbpedia.org/ontology/' shortened to 'dbo:' and 'http://dbpedia.org/property/' to 'dbp:'

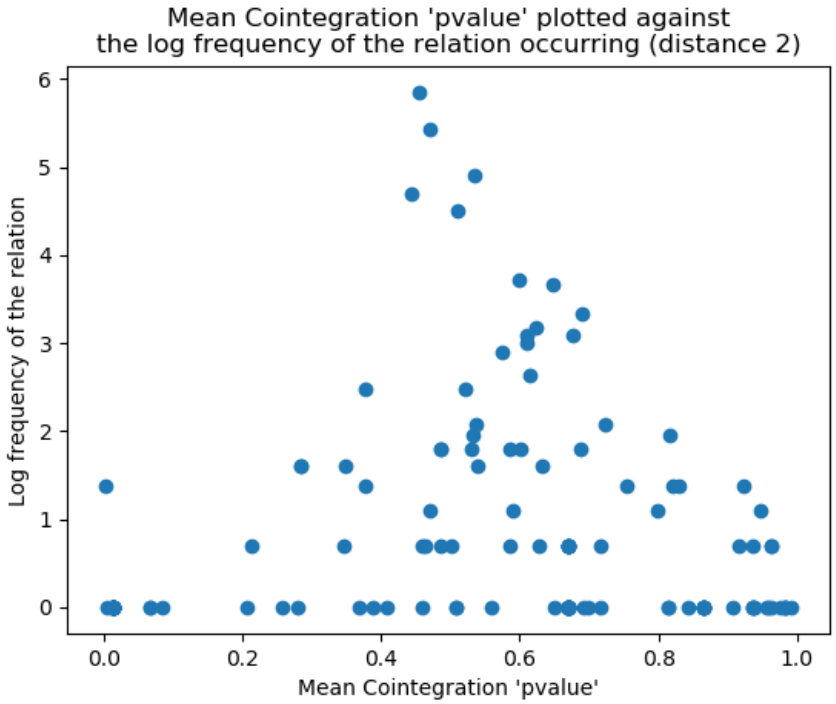

Figure 5.3: The log frequency of a distance 2 relation occurring plotted against the mean pvalue of pairs of companies which hold that relation

to construct SPARQL queries to detect the presence of these relations in pairs of the test symbols.

Using the set of test symbols none of the significant relations could be found. Whilst these results are disappointing it is not entirely surprising considering that when initially identifying the significant relations the highest frequency occurrence of any significant relation was only 4 with all the others occurring once. Therefore, expecting these relations which occurred so infrequently in the initial phase of this investigation to exist in the testing pairs may be considered somewhat unlikely.

6 Evaluation

6.1 URI Symbol Pairs matching

The automated system for identifying pairs of stock symbols and their DBpedia page was very ineffective. It was only able to identify 35 pairs out of 3434 symbols it checked and 2 of these were incorrect when checked manually. The most significant problem with this system was the lack of data available in the Web Ontologies. Many companies that appear on the NASDAQ exchange do not have a DBpedia page setup and therefore its simply not possible to identify these pairs, as they do not exist. Furthermore, of the companies that do have DBpedia pages many of them do not have their stock symbol stored on the page thus making it unfeasible to automatically identify these pairs.

By checking for the DBpedia pages to have additional relations, such as:

```
<{companyURI}> <http://dbpedia.org/property/tradedAs>
    <http://dbpedia.org/resource/NASDAQ-100> .
```

It may be possible to reduce the amount of stocks misidentified. However due to there being so much missing data in the ontology adding this constraint would likely lead to even less pairs of URIs and symbols being found leading to more needing to be added manually anyway.

Whilst some companies URI symbol pairs were added to the list of pairs manually due to the volume of companies traded it was not feasible to find every pair with a DBpedia page and add it.

6.2 Data Collection and Representation

The Alpha Vantage API provided an excellent method for collecting stock data. It was easy to use and enabled us to collect all the data we required for free.

Storing time series data in turtle format is somewhat uncommon however by using this representation it was incredibly simple to construct SPARQL queries to analyse and extract data from it.

6.3 Detecting Relations In The Ontology

Of the companies that did have DBpedia pages the information stored within them varied wildly. Therefore due to this missing information it was very difficult to calculate a coefficient to measure the likelihood of the two being cointegrated regardless of what scoring method was used, i.e. checking for the existence of specific relations or the number of length n paths between the pages.

From our original Motivation we stated that both Qualcomm and Intel are members of the semi-conductor industry. In DBpedia the relation exists:

```
dbo:Category:Semiconductor_companies <http://purl.org/dc
   /terms/subject> dbr:Intel .
```

But the equivalent relation for Qualcomm:

```
dbo:Category:Semiconductor_companies <http://purl.org/dc
   /terms/subject> dbr:Qualcomm .
```

Does not exist. Had this relation existed within the Ontology this could have helped to identify a significant relationship as Intel and Qualcomm were found to have a significant cointegration relationship. It is this type of missing information within the ontology that may be one of the key reasons for these methods having failed to yield good results. This lack of information will have affected both methods attempted.

7 Conclusion and Further Work

Over the course of this project we have developed a hypothesis and a method for testing it. From our results we are unable to reject our null hypothesis as no significant link between the level of cointegration between a pair of stocks and information stored within web ontologies could be found, using either of the methods.

Whilst we were unable to draw a conclusion that this link exists this project has still provided a valuable contribution to the field by producing a set of methods that in the future could used to identify this link. More specifically, further work could utilise a larger set of stock symbols and most importantly have access to a much more data rich ontology which by then applying these methods a significant link between cointegration and company information could be discovered. From our investigations no current ontology is available for public use at this time, therefore, this would also require constructing one which would be a substantial project in its own right.

Considering the amount of companies traded on stock exchanges it would be a huge effort to manually construct this ontology manually. Liu et al. [13] present the concept of using Deep Learning to extract the semantic meaning of financial news headlines which can then be used to construct an ontology. In future work it may be possible to apply techniques, like this, in order to create a more data rich ontology to work from, which could lead to a significant link being found.

If a significant link were to be found, eventually, this then opens the avenue to further research where one could consider using the information stored in ontologies alone to help select candidates for Pairs Trading.

8 Appendix

8.1 Stock Symbol URI Pairs

Company	DBpedia URI	Stock Symbol
AudioCodes	dbr:AudioCodes	AUDC
Microsoft	dbr:Microsoft	MSFT
Aprotinin	dbr:Aprotinin	PTI
Innovative Solutions & Support	dbr:Innovative_Solutions_&_Support	ISSC
EBay	dbr:EBay	EBAY
Sykes Enterprises	dbr:Sykes_Enterprises	SYKE
Vodafone	dbr:Vodafone	VOD
Costco	dbr:Costco	COST
Intel	dbr:Intel	INTC
Ocean Power Technologies	dbr:Ocean_Power_Technologies	OPTT
Compugen	dbr:Compugen_(Israeli_company)	CGEN
Alphabet	dbr:Alphabet_Inc.	GOOG
Caesarstone Ltd	dbr:Caesarstone_Sdot-Yam	CSTE
Electronic Arts	dbr:Electronic_Arts	EA
Comcast	dbr:Comcast	CMCSA
Seagate Technology	dbr:Seagate_Technology	STX
Apple	dbr:Apple_Inc.	AAPL
Qualcomm	dbr:Qualcomm	QCOM
One Horizon Group	dbr:One_Horizon_Group	OHGI
Novavax	dbr:Novavax	NVAX
Support.com	dbr:Support.com	SPRT
PepsiCo	dbr:PepsiCo	PEP
Alaska Communications	dbr:Alaska_Communications	ALSK
Vistaprint	dbr:Vistaprint	CMPR
Alphabet	dbr:Alphabet_Inc.	GOOGL
Cognizant	dbr:Cognizant	CTSH
Bio-Techne	dbr:Bio-Techne	TECH
Cerner	dbr:Cerner	CERN
Amazon	dbr:Amazon.com	AMZN
Facebook	dbr:Facebook	FB
Texas Instruments	dbr:Texas_Instruments	TXN
Cisco Systems	dbr:Cisco_Systems	CSCO

Table 8.1: Automatically identified URI Stock symbol pairs - symbolUriPairsManual.ttl

Company	DBpedia URI	Stock Symbol
Nvidia	dbr:Nvidia	NVDA
Tesla	dbr:Tesla	TSLA
Activision Blizzard	dbr:Activision_Blizzard	ATVI
Formula Systems	dbr:Formula_Systems	FORTY
Sinclair Broadcast Group	dbr:Sinclair_Broadcast_Group	SBGI
Autodesk	dbr:Autodesk	ADSK
Zynga	dbr:Zynga	ZNGA
Western Digital	dbr:Western_Digital	WDC
Advanced Micro Devices	dbr:Advanced_Micro_Devices	AMD
Adobe	dbr:Adobe	ADBE
Sapiens International Corporation	dbr:Sapiens_International_Corporation	SPNS
Dropbox	dbr:Dropbox	DBX
Qorvo	dbr:Qorvo	QRVO
Trip Advisor	dbr:TripAdvisor	TRIP
Scholastic Corporation	dbr:Scholastic_Corporation	SCHL

Table 8.2: Manually identified URI Stock symbol pairs - symbolUriPairsManual.ttl

8.2 Method 2: Identification and Test Sets

8.2.1 Identification Set Symbols

NVAX, ISSC, INTC, SBGI, NVDA, EBAY, FB, EA, SPRT, ADSK, ZNGA, PEP, COST, TSLA, SPNS, CSTE, AAPL, PTI, CERN, MSFT, VOD, CGEN, GOOGL, QRVO, ADBE, CMCSA, TXN, TRIP, CMPR, AUDC, CTSH, ALSK, ATVI, DBX and CSCO.

8.2.2 Test Set Symbols

FORTY, AMD, STX, TECH, AMZN, OHGI, SYKE, WDC, SCHL, OPTT and QCOM.

Bibliography

[1] (2019). Alpha vantage, [Online]. Available: https://www.alphavantage.co/ (visited on 12/03/2019).

[2] (2019). Nasdaq companies, [Online]. Available: https://www.nasdaq.com/screening/companies-by-industry.aspx?exchange=NASDAQ (visited on 01/02/2019).

[3] C. J. Granger, 'Developments in the study of cointegrated economic variables', *Oxford Bulletin of economics and statistics*, vol. 48, no. 3, pp. 213–228, 1986.

[4] H.-L. Han and M. Ogaki, 'Consumption, income and cointegration', *International Review of Economics & Finance*, vol. 6, pp. 107–117, 1997.

[5] M. Whistler, *Trading pairs: capturing profits and hedging risk with statistical arbitrage strategies*. John Wiley & Sons, 2004, vol. 216.

[6] 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.

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

[8] P. McSharry, 'Efficient pair selection for pair-trading strategies', 2015.

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

[10] J. P. McCrae. (2019). The linked open data cloud, [Online]. Available: https://lod-cloud.net/ (visited on 30/01/2019).

[11] H. Qu, M. Sardelich Nascimento, N. N. Qomariyah and D. L. Kazakov, 'Integrating time series with social media data in an ontology for the modelling of extreme financial events', in *LREC 2016 Proceedings*, European Language Resources Association (ELRA), 2016, pp. 57–63.

[12] B. Bozica, J. Peters-Andersa and G. Schimak, 'Filtering of semantically enriched environmental time series', 2012.

[13] 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.

[14] R. Hotten, 'Volkswagen: The scandal explained', *BBC News*, 10th Dec. 2015. [Online]. Available: https://www.bbc.co.uk/news/business-34324772 (visited on 27/04/2019).

[15] B. Snyder and S. Jones, 'Here's a timeline of volkswagen's tanking stock price', *Fortune*, 23rd Aug. 2015. [Online]. Available: http://fortune.com/2015/09/23/volkswagen-stock-drop/ (visited on 27/04/2019).

[16] (2019). Volkswagen group brands models, [Online]. Available: https://www.volkswagenag.com/en/brands-and-models.html (visited on 28/04/2019).

[17] (2019). Company list (nasdaq, nyse, amex), [Online]. Available: https://www.nasdaq.com/screening/company-list.aspx (visited on 12/03/2019).

[18] J. Perktold, S. Seabold and J. Taylor. (2017). Statsmodels.tsa.stattools.coint, [Online]. Available: https://www.statsmodels.org/stable/generated/statsmodels.tsa.stattools.coint.html (visited on 12/03/2019).

[19] (2014). Scipy.stats.https://docs.scipy.org/doc/scipy-0.15.1/reference/generated/scipy.stats.spearmanr.html, [Online]. Available: https://docs.scipy.org/doc/scipy-0.15.1/reference/generated/scipy.stats.spearmanr.html (visited on 26/04/2019).