An Algorithmic Investigation of Conviction Narrative Theory: Applications in Business, Finance and Economics

A DISSERTATION PRESENTED
BY
RICKARD NYMAN
TO
THE DEPARTMENT OF COMPUTER SCIENCE

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY
IN THE SUBJECT OF
FINANCIAL COMPUTING AND ANALYTICS

UNIVERSITY COLLEGE LONDON
LONDON, UK

DECEMBER 2015

An Algorithmic Investigation of Conviction Narrative Theory: Applications in Business, Finance and Economics

Abstract

The thesis aims to make *conviction narrative theory* (CNT) operational and test its validity via a combination of text analysis, network analysis and machine learning techniques. CNT is a theory of decision-making asserting that, when faced with uncertainty, agents are able to act by constructing narratives that yield conviction. The developed methodology is *directed* by CNT and therefore limits problems related to spurious correlations frequently encountered in studies using large datasets. The thesis provides empirical support of the theory and how it can be used to understand the economy and financial markets. The thesis develops a *relative sentiment shift* (RSS) methodology that captures emotional variables within text archives and also tests the extent to which these can be accurately measured, to establish causal economic and financial relationships hypothesised by the theory to exist on a macro level. Better-than-economic-consensus forecasts of the Michigan Consumer Confidence index, statistically significant explanatory power of real US GDP growth, evidence of causality from relative sentiment to the most widely used measure of financial market volatility, the VIX, are obtained in the process. On the micro level, the RSS methodology is applied to particular narratives to test theoretical expectations showing how it can be used to measure the emergence of *phantastic object narratives*, narratives for which anxiety substantially disappears despite the existence of conflicting evidence. To illustrate the importance of the overall ecology of narratives to understand shifts in macro sentiment and financial stability, as well as a means to qualitatively understand the relation between the macro and the micro approach, a form of dynamic content network analysis is applied. Using the narrative model, measures of the degree of formation of a dominant narrative are shown to correlate with RSS and Granger-cause indicators of financial stability, such as the VIX and the S&P 500 index.

Contents

Listing of figures

List of Tables

Acronyms and Symbols

CNT	Conviction Narrative Theory
DNF	Dominant Narrative Formation
RSS	Relative Sentiment Shift
RTRS	Reuters News Archive
BROKER	Broker Circulars Archive
MCDAILY	Daily Market Commentary Archive
$\in$	Set membership
$\subseteq, \subset$	Set inclusion
$\wedge$	Logical conjunction
$\mathbb{Z}, \mathbb{Z}^n$	The set of integers and n-dimensional vectors of integers
$\mathbb{N}, \mathbb{N}^n$	The set of natural numbers and n-dimensional vectors of natural numbers
$\mathbb{R}, \mathbb{R}^n$	The set of real numbers and n-dimensional vectors of real numbers
$[a, b]$	Inclusive range of values from a to b
$\|S\|$	The cardinality or size of set S
x	A scalar or vector
$\hat{x}$	An estimation of the scalar or vector x
$\hat{M}$	An estimation of the matrix M
Δ	Difference operator applied to an ordered sequence of values
$\sum_i^j$	The discrete sum from index i to index j
$\int_i^j$	The integral from index i to index j
M^T	The transpose of matrix M
$tr(M)$	The trace of matrix M
$Distrib(\alpha)$	Probability distribution Distrib with parameter vector α
$p(X)$	The probability mass function of discrete random variable X
$E[X]$	The expectation or expected value of random variable X
$Var[X]$	The variance of random variable X
$Y\|X$	The random variable Y conditioned on the random variable X
$\|\|x\|\|^p$	The p-norm of the variable x

This thesis is dedicated to my family, friends and colleagues who have always shown the greatest support

Acknowledgments

I AM SINCERELY GRATEFUL to my doctoral advisors Dr Robert Elliot Smith and Professor David Tuckett, as well as to Professor Paul Ormerod, for their invaluable guidance throughout this work. Without their support and our many intellectually stimulating conversations this thesis would not have been possible. Further thanks are also due to Professor Philip Treleaven, Yonita Carter and colleagues at the Centre for Doctoral Training in Financial Computing and Analytics at University College London for their support over the years.

When the crisis came, the serious limitations of existing economic and financial models immediately became apparent. Arbitrage broke down in many market segments, as markets froze and market participants were gripped by panic. Macro models failed to predict the crisis and seemed incapable of explaining what was happening to the economy in a convincing manner. As a policy-maker during the crisis, I found the available models of limited help. In fact, I would go further: in the face of the crisis, we felt abandoned by conventional tools.

Jean-Claude Trichet

1 Introduction

1.1 Motivation

Since the onset of the global financial crisis and subsequent economic slowdown in 2008 there has been a growing recognition of the need for scientific reform underpinning public policy, most notably in the field of economics. When faced with the need for a response to the crisis, policy-makers were left dumbfounded by standard economic theories and tools. The governor of the European Central Bank at the time, Jean-Claude Trichet, summarised the situation in the leading quote of this Chapter [110]. What are the conventional tools which Trichet referred to, and why did he feel abandoned by them?

Although this thesis will only briefly reference economic theory, and is best considered as purely complementary to economics, it is nonetheless important to have a general understanding of some of the most important themes of economics to see how the contributions of this dissertation fit into academic and policy discourse. This dissertation primarily grew out of a need for empirical support of a new theory of decision-making for science and policy, in order to understand how the 2008 crisis came about so that similar crises in the future can be mitigated.

Two central themes underpinning mainstream classic economic theory are *rational choice theory* and *general equilibrium theory*.

Rational choice theory essentially asserts that agents can be modelled as having

fixed preferences and therefore make 'consistent' decisions. This assumption, although often criticised for oversimplifying real human behaviour, enables mathematical models of agents in a social system. Such models can provide policy-makers with simple answers to very challenging questions, and are therefore important. However, an implicit assumption of rational choice theory is that there is no real *uncertainty*, as far as the agent is concerned, about the future state of the world. Given a set of possible choices, the agent can always apply some form of cost-benefit analysis to arrive at a correct answer. Furthermore, assuming the world remains the same, the agent will consistently make the same decision when faced with the same alternatives. The world can therefore be modelled *as if* it was completely deterministic.

General equilibrium theory attempts to explain the relationship between supply and demand in the economy. It seeks to prove that a set of prices exists that will result in a 'general' equilibrium. One aspect of the theory aims to determine under what conditions the economy will be in such an equilibrium. Thus, all that follows from a theory built on this assumption will be consequences of an already assumed price equilibrium.

Within classic economics there is therefore no place for the role of uncertainty and market prices are assumed to be in a state of equilibrium. Because it was implicitly assumed that the world could be treated as deterministic, there was no real need to consider *human and social psychology*. The field of economics achieved analytic tractability and the intellectual prestige that tends to go with it. However, given this intellectual framework of mainstream economics, and its semantics, it becomes very difficult to explain, and anticipate, why preferences might suddenly change in such a way that market prices appear out of equilibrium. Indeed, a research agenda founded on these assumptions cannot explain economic recessions or financial crises, other than calling them temporary fluctuations away from the assumed equilibrium caused by random external shocks. The tools derived from mainstream economics could therefore not prepare policy-makers for the sudden 'shock' that, based on standard economic and financial measures, was completely unforeseen. Neither could economic theories help explain what happened *ex post*.

Of course, there are notable exceptions among economists, such as Nobel price winners George Akerlof and Robert Shiller, who have encouraged the community to break free from some of these assumptions. It is important to note that many of the assumptions made, although not realistic in many ways, have been hugely helpful in formulating models and making it possible to understand some features of the economy and to make some predictions. However, the states of the economy that are arguably the most important to understand and predict, given their enormous consequences for society, are the 'unstable' ones - those that are conceivably *out of*

equilibrium. Consequentially, these features of the economy are left impossible to understand and predict.

> "We need to develop complementary tools to improve the robustness of our overall framework. In this context, I would very much welcome inspiration from other disciplines: physics, engineering, psychology, biology."
>
> Jean-Claude Trichet

The work of this thesis follows the advice of Trichet of the need for a multidisciplinary approach to understand what drives the economy. There is a need for tools and techniques from outside classical economics if similar crises are to be avoided in the future, or at least the negative impact to be limited. For example, it is known from mathematics, in particular the field of chaos theory, that even seemingly simple dynamical systems, of which the global economy can be considered an immensely complicated one, can appear entirely unpredictable. In other words, such a system can behave *as if* it was totally random. If it is acknowledged that the world can in some cases be thought of as unpredictable, it follows that agents in the system cannot rely on optimisations under constraints when making decisions. There will be important decisions to be made for which the agent cannot know, a priori, using mathematical tools relying on historical statistics, what the optimal choice might be. In this setting, the agent's mental state must play a significant role in determining the outcomes of the system. It is therefore necessary to understand and to measure this 'mental state'.

A new theory of decision-making under uncertainty is being developed at the Centre for the Study of Decision-Making Uncertainty at University College London (UCL). *Conviction narrative theory* (CNT) states that agents act under uncertainty by forming narratives that induce conviction - a predominance of approach emotions over avoidance emotions (i.e., the degree of conviction can be measured by the relative balance between these two emotional groups). The theory initially grew out of interview data with senior fund managers around the world [111] conducted by David Tuckett. The aim of the study was to understand how fund managers are able to invest billions of dollars on a daily basis while faced with real uncertainty about the future outcomes of those investments. CNT emerged as a unique approach for understanding decision-making under these conditions. It therefore fits the role of an alternative to rational choice theory that applies to those situations where agents cannot know a priori what the optimal choice is.

New computer science methodologies and digital data sources have recently made it possible to study the effects of mental states, both for the individual agent and the

society as a whole, in more rigorous and analytic ways than previously possible. A unique opportunity therefore exists to empirically test conviction narrative theory as a general approach to understand decision-making, and more specifically how it can be used to explain economic recessions and financial crises without resorting to labelling them 'random shocks'. This is the main motivation of this dissertation and what it aims to achieve.

1.2 Research Objectives and Methodology

This section outlines the main objectives of the thesis to support its hypothesis. The overarching hypothesis can be formulated as follows.

1.2.1 Hypothesis

> This thesis examines the hypothesis that CNT can be made operational and aspects of it empirically tested via a combination of text analysis, network analysis and machine learning techniques. In particular, three hypotheses relating to CNT will be tested. One hypothesis of CNT tested is the potential predictive relationship between macro confidence (as defined by the aggregate emotional conviction of a social group, such as the participants of an economy) and economic growth. A further hypothesis of CNT tested is the potential to use the derived methodology to detect the phenomena known as *divided-state* mentality with respect to a given conviction narrative, represented by a disconnect between conviction and reality for a given social group (where the group dynamic is referred to as *groupfeel*), potentially causing financial bubbles (i.e., *causing* some of the 'shocks' of classic economic theory). The final hypothesis tested is that those narratives that emerge as *dominant* narratives of a social group have a predominance of either approach or avoidance emotion (thus, emergence of dominant narratives correlates with shifts in macro confidence) and times when dominant narratives are formed could therefore have a negative impact on financial stability (which relies on heterogeneity of views).

The hypotheses can be more clearly defined. There are two main hypotheses postulated by CNT.

1. It is hypothesised that a relative increase in the dominance of approach or avoidance emotions (loosely defined as *excitement* and *anxiety*) in overall

economic narratives (i.e., from a macro point of view, in a country, sector, region) indicate changes in economic confidence and so actions that impact the economy and economic growth.

2. It is hypothesised that a significant relative gap in the dominance of excitement over anxiety around particular topics (i.e., from a micro point of view) indicate that thinking and decision-making around these topics have become unbalanced, creating a "divided state" - suggesting the build-up of systemic risk.

A further hypothesis is postulated in this thesis as follows.

3. It is hypothesised that a significant shift in macro confidence occurs exactly when a dominant narrative has emerged. In other words, the connection between macro confidence and micro confidence is such that a narrative becomes dominant whenever there is a large shift in macro confidence. Furthermore, the formation of such a dominant narrative is detrimental to financial stability, as financial markets rely on heterogeneity of views.

To test these three hypotheses three objectives are defined, which when met will serve as tests of the claims in the thesis statement. The objectives will be dealt with as three distinct experiments.

1.2.2 First Objective; Testing CNT with a 'Macro Analysis'

By developing a sentiment analysis methodology (called *relative sentiment shifts*, or RSS), directed by CNT and applied to all narratives in a given text collection (i.e., a 'macro' analysis), the relevance of CNT to the understanding of economics and finance can be empirically tested. The objective is to test relationships hypothesised by CNT [112] between key elements of conviction narratives and economic and financial variables as stated in the first hypothesis.

The first experiment applies the sentiment analysis methodology developed in Chapter 4 to a range of different data sources and shows that the metrics derived consistently yield significant predictions of relevant economic and financial variables as expected by CNT [112]. As such, the idea that these features of CNT can explain economic and financial variables, as hypothesised, is tested. It is particularly believed by CNT that the degree of confidence in the economy has a causal relationship with economic growth [115].

1.2.3 Second Objective; Testing CNT with a 'Micro Analysis'

By focusing the sentiment analysis methodology on individual case-studies (i.e., a 'micro' analysis), in the form of particular narratives selected ex-post as descriptive

pathological cases, the second objective aims to support the belief that the central concepts of CNT known as *groupfeel* and *divided mental states* can be measured in particular narratives (so called *phantastic object narratives*, or simply *phantastic objects*). It is argued by CNT [114] that the existence of these features are key characteristic drivers of 'irrational exuberance' as stated in the second hypothesis.

In the second experiment the same methodology as used in the first experiment is applied to a set of case-studies, ex-post argued to be pathological cases of conviction narratives subject to the central themes of *groupfeel* and *divided states*. Thus, the presence of these central concepts of CNT can be detected empirically.

1.2.4 Third Objective; Relating the 'Micro' with the 'Macro' Analysis

By developing a form of narrative content analysis to relate the 'macro' sentiment with the collection of 'micro' sentiment that together make up the macro, the third objective aims to qualitatively explain and understand the macro sentiment measure to uncover the clusters of 'entities' most related to macro shifts. This further strengthens the support for the first objective, and makes feasible the task of finding narratives and entities ex-ante that could negatively impact a business, an economy and/or the stability of the financial sector. Furthermore, the relationship between narrative homogeneity and shifts in RSS as well as financial stability can be tested, as stated in the third hypothesis.

In the third experiment a form of content analysis that relates narrative 'entities' and their individual connections with the macro sentiment measure is developed. This analysis provides a procedure by which to test the third hypothesis.

1.3 Research Contributions

Due to the interdisciplinary nature of the thesis, its contributions span several scientific fields; social-psychology, economics, finance and computer science. The main contribution is summarised in the following paragraph.

> The primary contribution of this PhD thesis is a theoretically justified and focused computational methodology to study a social-psychological theory (CNT) and make hypotheses derived from this theory empirically testable. Furthermore, the results from the hypotheses that have been tested in the various experiments support the validity of the theory.

In social-psychology, economics and finance, the thesis serves as evidence supporting CNT as a theory of decision-making and its consequences for economics,

finance and risk. The thesis supports a central hypothesis of CNT about the significant role of narrative and emotion to the state of the economy and financial markets and might help detect narratives subject to unstable mental and social processes - in other words, the methodology can potentially be used in the area of 'cognitive and social' risk monitoring. The main results are published in [114], [115], [78] and [77] and have been presented at numerous conferences.

In computer science, the thesis expands, in a general sense, on the discussion and understanding of large unstructured data sources and the ability and limitations to extract valid scientific relationships from them. Literature concerning the accuracy of 'traditional' scientific hypotheses is related to those formed by analysis of so called *big data*, thus viewing new and existing big data methodologies from a different (but 'old') perspective. The thesis argues how best to tackle these issues by developing theoretically justified algorithmic methodologies, and also serves as an example of such an approach. The argument has been presented at a conference at the European Commission in Brussels [81].

More specifically, the contributions made in the thesis can be listed as follows.

1.3.1 Contribution 1: A Methodology for Lexicon Evaluation in Time Series Sentiment Analysis

Chapter 4 derives a methodology to measure the relative balance between approach (excitement) and avoidance (anxiety) emotions. Having first argued for a word-lexicon approach to time series sentiment analysis (for the purpose of operationalising CNT) as opposed to a machine-learning classification approach, several criteria to measure the accuracy of a given lexicon are defined. This facilitates the evaluation of alternative lexicons for the purpose of capturing latent emotional time series variables. The methodology for evaluating this property of a given lexicon and for adjusting the lexicon based on these criteria is the first contribution of the thesis.

1.3.2 Contribution 2: Predicting GDP and MCI using RSS

Chapter 5 (experiment one) applies the RSS methodology developed in Chapter 4 to test the first hypothesis listed above, i.e., if changes in economic confidence impact the economy. It is found that the confidence index derived for the US economy through an analysis of Reuters news articles published in New York and Washington can be used as a significant explanatory variable in a regression of US GPD growth, even when controlling for other forward-looking variables such as asset prices (and other measures of confidence and anxiety such as the Michigan Consumer Sentiment Index and the VIX, which are shown to be Granger-caused by RSS indices extracted from

other data sources and/or Reuters). This provides significant support for the first hypothesis. Furthermore, it shows the potential to forecast, as well as nowcast[1], GDP growth using the RSS methodology. This is the second contribution of the thesis.

1.3.3 Contribution 3: Operationalising Phantastic Objects, Divided States & Groupfeel using RSS

Chapter 6 (experiment two) applies the RSS methodology to particular narratives (topics) hypothesised to show characteristics of conviction narratives believed to be held in an unbalanced, 'exuberant', state of mind. It is found that stories about Fannie Mae (published in Reuters) exhibited a significant decline in anxiety relative to excitement over the period 2005 through 2007, despite the fact that housing prices started to decline almost 18 months before the end of this trend in RSS. The pattern is consistent with expectations from CNT regarding phantastic object narratives. Similar relative sentiment patterns are observed when the analysis is focused on Enron emails regarding two specific business projects. It is further shown that a key social group experienced significantly different sentiment patterns concerning these two business projects relative to the rest of the network. The group became substantially more anxious in the second half of the period, evidence of a divided state collapsing within this particular group. The experiment provides support for the second hypothesis. It also suggests that it is possible to distinguish the emotional patterns between different social groups, and that social context may therefore play a significant role in determining an individual agent's emotional state. In other words, supporting the idea that text analysis can be used to measure groupfeel-type phenomena. This is the third contribution of the thesis.

1.3.4 Contribution 4: A Methodology to Measure DNF and Establishing its Relationship with RSS and Financial Stability

Chapter 7 (experiment three) develops a new methodology to relate entities (in the context of the third experiment, the entities considered are organisations) within narratives to one another and the overall structure of the narrative landscape to RSS. It is shown that shifts in RSS correlates with the emergence of a dominant narrative (as measured by the methodology presented in Chapter 7). It is further shown that these new indices of *dominant narrative formation* (DNF) carry predictive information relating to financial stability, in the form of the VIX and the S&P 500 index. Thus, the results provide support for the third hypothesis. This is the fourth and final contribution of the thesis.

[1]The methodology was used in [77] to nowcast (predict the current value of) US GDP growth.

1.4 Thesis Outline

The thesis is structured as follows.

Chapter 2 presents the relevant theoretical (first section) and technical (second section) literature touched upon in the introduction. In particular, a brief section is devoted to what has been the dominant theory of decision-making in economics and finance, *rational choice theory*. This leads to a discussion and summary of a new theory of decision-making, CNT, which the thesis aims to make operational and empirically test. This is briefly followed by a discussion about the promises and pitfalls of new data and technology. These new tools can dramatically help to study the effects of human behaviour in many aspects of society, but can at the same time cause a lot of concern about the interpretation and validity of such found effects. This section sets the second theme of the thesis (CNT being considered the first); the importance of directing a search for patterns by theoretical expectations. The second part of Chapter 2 covers the technical tools in text analysis, network analysis and machine learning necessary to conduct the experiments. Chapter 3 presents the data used to test the hypotheses. Chapter 4 develops and evaluates the RSS methodology, applied in the first two experiments, from the point of view of CNT. Experiments 1, 2 & 3 are presented in Chapters 5, 6 & 7 respectively. Chapter 8 assesses the experiments, concludes the thesis and provides a summary of future potential research streams to pursue.

2

Literature Review

This chapter presents relevant literature, both theoretical and technical. The chapter is divided into two main sections. The first section deals with theory; in particular it deals with various theories of decision-making. First, it deals with the prevalent theory of decision-making adopted in many social sciences and most prominently in the field of economics, *rational choice theory*. Secondly, it deals with an alternative theory of decision-making with a socio-psychological background, *conviction narrative theory*. Finally, the first section ends with a more general discussion about the role of theory as we enter into what might aptly be called the *Big Data era*.

The second section of the chapter presents technical tools used to explore the thesis hypothesis: text analysis, graph theory, machine learning and statistics.

2.1 Theory Review

The world, as experienced by an intelligent agent, might be understood by the agent through simulations of alternative possible states. This is perhaps most notably apparent given the pleasure commonly derived from fiction. The world of social interactions is inherently too complex for the mind to fully comprehend and predict, and as such fiction is relied upon as approximated simulations of this interaction [68]. Computer simulations may be conceived of as the scientific equivalent of fiction. It should be recognised that, within the social sciences, 'proof by computation' (i.e.,

repeated simulation) and 'proof by analysis' (i.e., a complete mathematical proof) are achieved through the same fictional activity, where the difference simply lies in the complexity of the simulated social interactions and derivation of a 'solution'. Proof by computation is not, a priori, inferior to proof by analysis. However, in either of the two deductive settings it is crucial to support all axioms through empirical findings. Any theory that aims to explain real-world phenomena must in some way be based on real-world observations. It is therefore necessary to begin the exploration of a new theory through an inductive empirical investigation that seeks to find and test patterns and behaviour that govern the system under study. In other words, it is necessary to support and build a theory by an inductive investigation that is guided by axioms and that will iteratively modify them. Whenever observed patterns modify the axioms, the new axioms must be investigated further and tested on previously unobserved data, or the significance of any statistical test will in general lack meaning. It is important to be mindful of these fundamental principles [63].

However, predictions based on empirical studies alone are limited by the current parameter values. Thus, if the parameters change the system will most likely appear unpredictable again. Government officials and policy-makers, whose job it is to often change the parameters for better social systems, require both empirical findings as well as theoretical models. Since the well-known Lucas critique was formulated in 1976 [66] economists have recognised and promoted the importance of deductive models based on structural micro-foundations. The critique states that any model of the economy based purely on relationships found using aggregated historical data cannot be used by policy-makers when deciding to implement a change in policy as that change will likely also change the data-generating process itself, rendering the effects of policy changes unpredictable. In a structural model parameters can be changed and different states of the system can be observed and anticipated. This is perhaps one explanation of the widespread adoption and use of economic models in general, and the rational-agent model in particular, among policy-makers. However, at the time classic economic theory was developed, relatively little to no access to computational resources beyond the human brain was available and very limited access to granular data. Strong simplifications were therefore necessary and useful. In modern days, technology and data are no longer serious limitations. On the one hand, it is now possible to empirically study phenomena previously not amenable to rigorous analysis because of access to new data. On the other hand, it is now possible to model more complex interactions by the use of new computer technology. The field of economics must now undergo nothing short of a revolution to bring back those real-world observations that were left out for the sake of analytic convenience.

A careful interplay between empirical analysis on the one hand and theoretical

modelling (whether mathematical, simulated or purely abstract) on the other, can be considered the scientific gold standard.

The rest of this section will briefly discuss rational choice theory, the highly influential and much-talked-about corner stone of economic theory. Following this, attention will be directed to an alternative, recently developed and less restrictive (in the sense of assumptions made), theory of decision-making known as *conviction narrative theory*. The final section discusses some of the potential pitfalls of big data analysis (as well as some of the benefits) and how it might be possible to avoid them.

2.1.1 Rational Choice Theory

Rational choice theory is a framework in which to understand and model social and economic behaviour, perhaps most notably used within the field of economics. Its origins can be traced as far back as to economist Adam Smith in his groundbreaking work *The Wealth of Nations* [99]. Within modern economics, the theory simply mandates that agents *consistently* rank choice alternatives by the use of some sort of cost-benefit analysis [47]. For example, if the agent prefers A to B and B to C, then the agent also prefers A to C. In other words, alternatives can be consistently ordered. Thus, agent preferences are assumed to be static, i.e., fixed over time. This facilitates the existence of an optimal choice. One consequence of this is that, all else being equal (i.e., the characteristics of the potential choices have not changed), the agent does not change to prefer B to A if A was previously preferred.

Two major commonly made assumptions are:

1. The agent has perfect information about the outcomes of any choice, or a probabilistic representation of those outcomes

2. The agent has the cognitive ability to weight every choice against every other by using this perfect information (theories of *bounded rationality* relax this assumption)

The proponents of the use of rational choice theory do not claim these assumptions to be a full description of reality, and it is in fact very easy to find real counterexamples. What they do uphold is that both good and bad models can aid in reasoning and help to formulate hypotheses, whether the models are intuitive or not [42].

2.1.2 Conviction Narrative Theory

The socio-psychological theory of decision-making known as Conviction Narrative Theory (CNT) is a theory aimed at understanding and explaining how agents make

decisions when 'optimal' choices cannot be made [112], [26]. Put differently, the theory discards the first major assumption about the rational agent, the existence of a fully predictable environment. The theory is based on the in-depth analysis of interviews in financial markets and theories hypothesised to explain the development and fragmentation of market confidence [111]. It aims to explain how agents are able to make decisions and act under such uncertainty at all, not to mention how they often manage to make good decisions. If the probability space of the outcomes of two given actions cannot be known, it is impossible to make a choice between them based on mathematical calculation alone. Consistency of actions can therefore not be ensured by mathematics and logic since there is no longer any reason to expect an agent to consistently make the same choice. Other factors must play a role, in particular the agent's environment (put simply, what other agents are choosing and other external factors) and mental state (put simply, how the agent feels about a given choice).

Outcomes of most real-world decisions of any significant interest, important to economic and financial life, cannot be described probabilistically. As a consequence of recognising this, no optimal choice can be selected based on a consistent ordering of derived utility from alternatives. If this is true, then how do agents single out a choice from equally plausible alternatives? CNT states that agents make decisions, and therefore act, by creating narratives (stories) that generate enough conviction about potential gains from the given action while at the same time suppress anxiety and doubt about potential loss. Agents must rely on *embodied simulations* of possible future outcomes to anticipate the future emotional response as experienced physically and mentally by the body (research in modern cognitive neuroscience suggests the physiological experience of the body can be just as apparent while imagining a reward as while receiving a reward [20]). Indeed, agents have adapted to such decision-making and the cognitive and emotional capacities have likely evolved to work together extremely effectively to conquer uncertainty without surrendering to inaction.

A conviction narrative can be thought of as an internal representation of an agent's environment that allows the agent to feel confident enough to make decisions under uncertainty. Such stories perform the narrative function of combining cognitive and emotional elements into a coherent temporal picture, and, by their nature, do not give due consideration to all possible scenarios. Ordinarily, commitment to a course of action requires agents to build their awareness of the potential advantages and disadvantages of a course of action, so that they come down on the side of the advantages. For this to be possible conviction narratives normally need to contain both attractor and doubt-defusing elements. The first makes the object intrinsically attractive, while the second manages any doubts that this might not be the case [112], [26].

The theory also hypothesises that conviction narratives always have the potential to become *phantastic object narratives*, that is to say, stories about supremely attractive objects (ideas or entities) of such unquestionable value that real-world facts potentially provoking doubts are ignored, as in some love affairs [113]. In this situation, termed a *divided state*, the world remains much the same, but stories about it have become (for a time) highly partial. A divided state is recognisable when attractive features in the narrative have grown and intensified in such an exciting way that the existence of doubt diminishes substantially or even vanishes from the narrative altogether [113]. Seen from a different perspective, one way to achieve the predominance of excitement over anxiety and doubt needed to commit to a course of action is for the brain to suppress anxiety-provoking information altogether. This type of divided state cannot, by its very nature of being disconnected from reality, be sustained and anxiety will therefore eventually come back with force. There could of course be solid grounds for less doubt if the world has changed accordingly, perhaps by some technological innovation, although it would be a cause for concern.

Psychologically, divided states are simplistic states in which the potentially anxiety-producing ideas that are ordinarily thrown up by experience (such as the possibility some bit of news is indicating an investment project may fail) are present in the mind, but not in such a way they become salient for conscious attention and thinking. Divided states are individual mental states, but in groups they can be supported institutionally by what can be called *groupfeel* - a state of affairs sometimes found in organisations where a group of people orient their thoughts and actions to each other based on a powerful and not fully conscious wish not to be different. In such group states members attend only to an "inside view" [54] and engage in groupthink [53]. Clearly, phantastic object narratives cannot maintain conviction forever. Eventually, because the underlying state of the world never actually changed, reality returns with force, the divided state gives way, and there is disappointment and revulsion at what was the phantastic object [113].

The phantastic object theory just summarised provides a potential explanation for first the evolution and then the collapse of financial market bubbles. The overall hypothesis that this set of ideas creates for testing is that conviction narratives in social networks infected by phantastic objects will:

- Increasingly exhibit a gradual elimination of doubt. It is this phenomenon which eventually produces a financial bubble or similar mass phenomena.
- Eventually exhibit a massive reintroduction of doubt, as the divided state and the phantastic object narrative collapse.

Phantastic objects, in the form of particular conviction narratives, may account for the emergence of financial market bubbles, but the overall degree of conviction of a social group (defined as 'confidence' for the sake of distinguishing it from the conviction of a single agent), or the economy as a whole, may account for shorter term fluctuations in the degree of business investment. As such, one important hypothesis of CNT is that the overall degree of confidence in an economy can have an impact on economic growth.

A further important aspect of the theory concerns how narratives spread in social groups to become dominant and how they might dissipate. Networks of information channels are believed to play a significant role in determining how narratives spread, with some individuals acting as more important than others [114].

The conviction of individual agents, the overall confidence of agents as a group, and the network dynamics influencing how narratives spread from the individual to the group are likely connected in complicated ways with sustainable economic and financial growth as well as unsustainable economic and financial growth. Towards the end of the thesis the concept of the formation of a dominant narrative will be introduced that might provide some insight into how these concepts could be related. In particular, it is found that dominant narratives emerge in a group exactly when there is a shift in confidence (i.e., conviction or the absence of it is oriented towards a particular narrative) and that this is a sign of instability.

The concept of a narrative in CNT is grounded in social-psychology, e.g., the work by Baumeister, Masicampo and Bruner [14], [24]. There are, of course, other definitions and formalisations of a narrative, e.g., mathematical abstractions [10]. This thesis will only focus on aspects of CNT.

2.1.3 The Big Data Age

The beginning of the 21st century may be considered the time when society entered into a new digital age. More information than ever before is made available in digital form. This information covers most aspects of our society, from private and confidential to public and open information. This section presents a somewhat informal discussion on some of the new methodological challenges facing meaningful statistical analysis of such information, which sets the scene for a significant theme of the thesis; the need for theory to guide the search for patterns in large emerging data.

Although these new large data sources are tremendously helpful and have revolutionised many areas of science, they also pose new methodological and ethical challenges. The existence of larger and more complete data sources, and algorithmic methods of analysing them, challenge the normative procedures of statistical

significance testing, e.g., [63]. This section attempts to clarify these issues and promote a philosophical framework for understanding, and safeguarding from, false conclusions.

With Big Data there is a fundamental challenge to determine validity and significance of established relationships - the concept of 'spurious correlations' that may pass traditional statistical tests with flying colours but fail to have any *structural meaning* whatsoever. A similar concept in machine learning is often referred to as 'over-fitting'. However, it can be argued that there is an important distinction. When talking about the potential for over-fitting a model to the data, it is already assumed that a particular relationship exists (the problem is to identify the nature of that relationship, i.e., a problem of model selection). Often the spurious-correlation problem arises when explanatory power and predictive accuracy are conflated [97]. By taking more seriously the predictive accuracy of explanatory models some spurious correlations might be avoided [97]. It is further argued in [119] that explanatory models (theories), particularly in the field of sociology, require predictability in order to be scientifically valid. Conversely, assuming that a predictive model has explanatory power (in the sense of providing a causal explanation) might lead to nonsensical theories [97]. In a New York Times bestselling book, the authors advocate that causal models are no longer necessary because the only thing needed for prediction is a statistical relationship [70]. This section argues that such an approach is a step backwards for science and that, in fact, causal theories are highly necessary (perhaps now more than ever). Further issues with large sample sizes are presented in [63], along with several concrete examples of what they consider to be questionable conclusions drawn from large samples. However, not all Big-Data research falls into this category. For example, a recent study into the predictability of private traits and attributes using Facebook data [58] used prediction accuracy as well as manual inspection of predictive features to evaluate the strength of found patterns.

Traditional means of statistical significance tests were conceived at a time when it was time and energy consuming to construct a scientific hypothesis. These tests were applicable only when the researcher tested a hard-earned hypothesis on previously unseen data. Every machine-generated model (a model with a specific set of parameters) is a scientific hypothesis. Therefore, an algorithm can generate millions of hypotheses in negligible time if applied to real Big Data. To see the implications, consider the following simple thought experiment:

- You have a hypothesis generator capable of generating 'semantically valid' random and independent hypotheses from the space of all hypotheses
- Imagine a conventional 95% significance level for these tests

- Imagine that 20 hypotheses are independently generated

Given these hypothetical assumptions, it is expected that at least one 'significant' relationship is found (i.e., achieves a p-value of less than 5%) since they are random by construction. Of course, it is easy to imagine that the probability that a random hypothesis is actually true is very small. If a further assumption is made that all positive findings are published and all negative findings are not, the conclusion would be that most scientific journals are largely incorrect. This has been extensively discussed in the literature, and is called a *positive publication bias* [31], [40]. The situation often comes about because the researcher puts aside negative findings due to a lack of interest (by the researcher, journals or the community at large), known as the *file drawer effect* [89]. The consequences for the research community as a whole, of neglecting negative findings, is a lack of awareness of what statistical tests have been performed by others in the past. If one team tests a given hypothesis and does not report the result, whatever it might be, other teams are more likely to test the same hypothesis. The above experiment applies in a modified form if twenty teams have tested the exact same hypothesis independently (on independent data samples). Therefore, a bias of positive publications will dramatically reduce the confidence one can have in any of the statistical tests performed. Techniques have been developed in the field of meta-analysis to detect publication bias and make corresponding adjustments (see for example the concept of funnel plots).

The situation for a given research team is luckily not as bad as the above thought experiment. Human-generated hypotheses are often not random and they tend not to be independent. In fact, some are even replications. If the hypotheses tested are expectations from the same scientific theory the probability of incorrectly judging something random as true is dramatically reduced. In other words, false positive findings are much less likely to occur if the tested hypotheses are somehow *dependent*. Thus, the main human resource is intuition, the existence of a *theoretical prior* belief. Hypotheses tend not to be random - the hypothesis 'pigs can fly' would probably not be tested (a human excludes a vast number of hypothesis without almost any mental effort, especially if guided by theory). However, algorithmic methods tend to be much less restrictive (the hypothesis 'pigs can fly' might actually be tested, and in fact a machine would test this or similar nonsensical statements unless explicitly told not to as specified by the form of the model and the constraints on its parameters). In this situation the researcher is much more likely prone to what is known as post-hoc theorising (or HARKing - Hypothesizing After the Results are Known) [55]. In other words, a relationship is found, an explanation for it is then formulated and the statistical test has been unknowingly performed on the data that generated the

Table 2.1.1: Contrasting hypothesis-testing methodologies; machine driven vs. theory driven

Test space features	Machine driven approach	Theory driven approach
Unrestricted feature space?	Yes, 'unlimited' number of tests and features on big data. 'Positive' results are often post-theorised	No, effort + time to perform a test. Features and hypotheses are filtered by theory. In other words, tests are not 'wasted'
Independent evaluation of results?	Yes, in most cases. The 'negative' results are discarded even though potentially relevant	No, one test result effects interpretation of others as they rely on the validity of the same theory

hypothesis in the first place, so called double dipping. This also happens implicitly if a model is respecified and tested on the same data set until the necessary statistical tests have all passed.

Furthermore, the space of possible hypotheses applied to Big Data is truly massive, it is therefore relatively easy to formulate plenty of nearly random hypotheses.

The two concepts of *independently-tested* and *nearly-unlimited* hypotheses, and the distinction between the conventional method of testing and the Big Data approach can be summarised in a simple two-by-two table, Table 2.1.1.

To summarise, with a brute-force algorithmic testing methodology, the researcher runs the risk of mistakenly ignoring negative findings that might in fact be related to a result later found to be significant (which would have questioned the significance of that result). With Big Data the space of possible tests to make is so vast that it is tempting to search in a much less restrictive manner than what has been the scientific convention. The convention of observing a small set of data to formulate a theory and a set of related hypotheses that can later be tested on unobserved data, has not properly carried over to the *Big Data era*. In the conventional situation, tests were much less likely to be 'wasted' on hypotheses that ex-ante were not considered deserving enough of serious attention, but which ex-post can easily be explained by post-hoc theorising.

Tackling Spurious Correlations

Hal Varian has done influential work on applying web-search-frequency data to predict, or 'nowcast' (that is, predict the present value of a variable), various economic time series [25]. Because web search is a relatively recent invention, the collected data do not reach very far back in time. For example, Google search

frequency time series only go as far back as 2004.[1] Therefore, to fit a model on economic data such as GDP growth, which is only available at a quarterly frequency, at the time of many of these studies less than 50 data points would have been available. On the other hand, the number of Google search categories or individual search terms to consider in a predictive model reaches the millions, if not billions. It is therefore trivial to find many combinations of explanatory variables that would fully explain the variance of the dependent variable, GDP growth. This problem is commonly referred to as *fat regression* [94]. Varian applied a combination of Bayesian methods to select the most likely explanatory variables [94]. The selected variables were then suitably averaged using a Bayesian model averaging technique. Although this will likely increase the robustness of the model's predictions, it will also likely obscure what is actually going on. As a decision-maker it might not be very satisfying to make decisions based on the result of a model average. However, the Bayesian approach does allow for specification of prior beliefs of the importance of any individual predictor, and in the most trivial case, full weight could be given to a particular predictor and zero weights to all others. This would reduce the approach to one directed purely by prior beliefs or expectations. In other words, in this limit it would be equivalent to a purely theory-driven approach.

A second approach to tackle spurious correlations, which may seem obvious and trivial at first but has strangely become unfashionable in Big Data analysis, is to formulate your hypothesis and methodology on the basis of an established scientific theory before variable and model selection. If instead of jumping into Big Data space with the aim of achieving a certain task such as predicting GDP growth, by any means necessary, one poses the question of what factors within the data might reasonably explain growth in GDP (and how) and perhaps even cause it, any predictive power then achieved will unlikely be spurious and might even predict ex-ante going forward in time. However, a certain amount of machine learning must likely be used in order to optimally benefit from the vast amount of data now available. It is likely that a combination of machine learning and variable and model selection based on a theoretical prior would yield superior results to using either approach in isolation. It was illustrated in [27] and [28], for different applications, that such a combined approach does indeed lead to state-of-the-art predictive performance.

Simply because much larger amounts of data now exist than ever before does not imply there is no longer a need to formulate hypotheses before they are tested. This approach is of course a special case of the above Bayesian framework in which full prior weight is given to a small set of variables pre-selected by a theory to be relevant

[1] At the time this thesis was written, Google provided a service for accessing the relative frequencies of different search terms used on Google at the web-address http://www.google.co.uk/trends/

in the given context. The great benefit of this approach is that it is much more easily communicated and relies entirely on *structural foundations* - a model for how the given system actually works and what variables should be treated as relevant.

This thesis exemplifies this approach by taking conviction narrative theory as the structural foundation and as such it will focus on very small set of variables expected to be relevant for decision-makers.

Bias-Variance Trade-off

The Bias-Variance trade-off is a statistical formulation of two competing sources of errors in generalisations of machine-learning algorithms [74]. The formulation applies to model selection and summarises the potential sources of errors a machine-learning algorithm may have when generalising to unseen data. The 'bias' term explains errors stemming from having selected a model space that is too narrow to be able to capture the true features of the training set, potentially resulting in *under-fitting* of the training data and therefore poor generalisations. The 'variance' term explains errors stemming from having selected a model space that is too large, potentially resulting in *over-fitting* of the training data and therefore poor generalisations. This formulation will be explained in more technical detail in the following section focusing on a technical review, but this statistical formulation will here be related to the above.

The issues discussed above can be restated in terms of this error composition. In the machine-driven approach there is a potentially massive space of models to fit to the training set (due to ease of testing), thus substantially increasing the risk of over-fitting. In the theory-driven approach the model space is restricted (biased) to a much smaller subset of potential models to explain the observed data, thus substantially reducing the risk of over-fitting (but with the potential risk of under-fitting). Owing to the inherent complexity of social systems, such as the macro economy, it is very unlikely that the true model will ever be found for any task at hand. Furthermore, the temptation to achieve short-term success in prediction will inevitably lure us into taking the machine-driven, less *model-constrained* approach. All this considered, the scientific community should always have a strong preference for the *biased* model-selection process provided by the theory-driven approach.

2.1.4 Narrative Text Analysis

What does the discussion of this chapter mean in practice? How can these different ideas be turned into a computational methodology used to test empirically the relevance of conviction narratives to decision-making under uncertainty?

As conviction narrative theory is applicable in a range of decision-contexts the

methodology must be as independent of data-type as possible. Thus, both the model defined and the features considered by the model should be as *invariant* over time, situations and social groups as possible. The success of the methodology will therefore be highly sensitive to the concept known as *over-fitting*. Given the trade-off to be made between accuracy on a given set of observed data and generalisability to unseen data, known as the Bias-Variance trade-off discussed earlier, it is therefore likely to be beneficial to *introduce* substantial bias to the methodology. The success of biased methodologies can be seen across many fields, for example when applied to financial asset portfolio-selection strategies [30], when used in the construction of decision trees for emergency care [46], or when attempting to understand the consequences of policy change [12]. In particular, a machine-learning approach to sentiment analysis would require a very large training corpus spanning several different data sources and contexts and also heavy regularisation. A sample from these data sources would have to be labelled by their degrees of excitement (approach) and anxiety (avoidance) so that relevant features can be learned. Such data are not readily available, and can often be very costly to compile.[2]

Furthermore, without a very large labelled data set such an approach is unlikely to be able to approximate human intuition of the emotional meaning of words (across contexts) to a reasonable accuracy. Words that are consistent with human intuition are likely to generalise better across contexts and datasources. *The methodology used will therefore rely on predefined lists of emotion words as features.* There are several such available lists. Hence, a key challenge becomes how to select the most suitable lists to use for the purpose of measuring the evolution of the approach and avoidance emotion groups, henceforth these variables will be referred to as excitement and anxiety respectively. This will be a major subject studied in depth in Chapter 4. Ideas about the *clarity* and *coherency* of lists of features as representations of the same underlying signal will be introduced and measured.

The selected single-word features can then be used to extract aggregate measures of excitement and anxiety from text data by a simple word-count methodology. However, even this very simplistic measure can be non-trivial to fully understand. The simplicity of the approach introduces substantial bias in our results which will likely improve the models ability to generalise to new text data sources. The simplicity also facilitates more computationally intensive stability and robustness tests.

For decision-makers in general and policy-makers in particular, the existence of a simple and transparent methodology is often a prerequisite for relying on that

[2]However, new crowd-sourcing platforms have recently been created to solve this problem (e.g., Amazon Mechanical Turk). Given the subtle nature of the features relevant to CNT, this approach would likely be more unreliable than the application of pre-defined lists of features.

methodology when making important decisions. An example of a big data application that serves to inform policy-makers at the highest levels of the European Commission is the European Media Monitoring system developed by the Joint Research Centre (JRC).[3] This system performs scheduled scrapes of news and other media sources from the internet in a range of languages, categorieses all incoming articles based on boolean pattern combinations and performs a variety of entity and event recognition tasks in a multilingual context. The JRC made a similar methodological decision as has been done in this thesis, to minimise model complexity as much as possible. This system is in use by several directorates of the commission as well as organisations such as the World Health Organisation, the African Union, the European Parliament and Europol.

Thus, a simple methodology does not only make it feasible to perform a variety of robustness checks as well as guarding against over-fitting your model, but is often required for explanatory purposes if the model is to be used in practice in any critical context. However, extensive research is focused on making machine learning models more interpretable [117], e.g., in medicine [61].

In Chapter 7 a text-analytic methodology is derived to test whether or not macro shifts in relative sentiment depend on increased homogeneity of the underlying narratives. If found to be the case, this would support a key aspect of CNT - that emotions are fundamentally generated through the construction of narratives, and thus spread socially via a 'narrative channel'. This methodology abstracts from much of the content and structure of individual narratives and focus instead on what can be considered to be the key entities talked about in news articles and whether or not they can be considered linked. This abstraction is made in order to understand, on a more intuitive level, the evolution of the links between narratives, as opposed to how the entities are being discussed more specifically. Given this level of abstraction, organisations (public and private) are considered to be the key entities within narratives expressed in news articles and a graph link structure is inferred between these entities from the underlying text database.

The material in Chapter 7 attempts to shed some light on questions such as:

- How contracted or dispersed is the 'space' of all narratives, and how does this dispersion evolve over time?
- How connected or related are the key entities mentioned within the narratives, and how does this connectedness evolve over time?

[3] At the time this thesis was written, the system was publicly accessible on the web address http://emm.newsbrief.eu/NewsBrief/clusteredition/en/latest.html

- When a shift in the macro relative sentiment series is observed, which are the entities within the narratives (or, more generally, which particular narratives) are most responsible for that shift? In other words, can the cause of a shift in RSS be detected?

- How does the narrative dispersion and connectedness measures relate to the issue of financial stability?

The aim of the exercise is two-fold. The first aim is to better understand the macro relative sentiment series and by doing so to, ultimately, extract more qualitative insights from it. Such explanatory insights would be useful to policy-makers using any type of sentiment indicator to inform their own judgement. The second aim is to relate the ideas of narrative homogeneity to the issue of financial stability. From the point of view of financial stability, a healthy financial sector relies on a heterogeneity of views, opinions and interpretations of data (i.e., heterogeneity of narratives). Without heterogeneity in markets, prices will unlikely clear at levels supporting long-term economic growth. If dominant interpretations of the future course of markets begin to emerge, this will necessarily manifest itself as more volatile and uncertain markets. Thus, a relationship between narrative heterogeneity and financial stability is tested for, which could help explain narratives' effect on financial instability.

The macro measure of relative sentiment constructed in Chapter 4 can be interpreted as measuring shifts in the degree of heterogeneity of emotion, a shift indicating less heterogeneity (more homogeneous emotion across the stories in the given database). Whether or not such homogeneity in emotion represents homogeneity of the underlying narratives that carries those emotions, the stories and interpretations of the world and the attention given to each, is not clear from the RSS measure itself. If it happens to be the case, that narrative homogeneity 'correlates' with emotional homogeneity, it might make it much easier for policy-makers to intervene and tackle the issue of the dominant narratives more directly, as they emerge. If, on the other hand, it turns out that macro shifts in emotion can not be attributed to specific dominant narratives the issue at hand would likely be much harder to address. The purpose of Chapter 7 is to shed more light on these questions.

CNT states that emotion is a key element of any narrative relating to beliefs about an uncertain future. Emotions are generated through the construction of a narrative, an interpretation of the world, and coexist with it. As macro emotions shift in a particular direction (become more homogeneous across discussions), whether towards relative excitement or anxiety, it could be expected that the underlying stories (from which these emotions are ultimately generated) would also be linked in some way. If this was not the case, then the emotions would in effect be generated *independently* of

the underlying narratives. The emotions would, if so, have spread via some other channel than their expression in narratives, perhaps through a 'biological' mechanism that does not depend on narratives at all (but simply on the basis of human interaction), or simply because of a statistical fluke. Such a fluke would occur if it so happens that a significant number of narratives share the same emotional pattern without in fact being related at all. Such occurrences would not be systematic (by their very nature of being random) so that simple statistical tests will tell if they have appeared or not.

Finally, beyond the scope of this thesis, the methodology derived can potentially be used to 'flag' particular narratives, or entities within those narratives, that show early signs of concern. In other words, early signs of becoming phantastic object narratives. If a suitable database contain narratives for which anxiety decays from a healthy and stable degree to levels no longer necessarily reflecting fundamentals (in other words, there are signs of divided state mentality), this methodology could be used to inform decision-makers that such narratives might have become phantastic objects. The entities analysed may in this case not be given by those studied here (organisations) but perhaps by other objects such as individuals or perhaps different financial assets. The methodology developed here can potentially be carried over to such applications given suitable sources of data.

2.2 Technical Review

This section presents some background on the technical tools required to understand the methodologies developed and the statistical tests performed in the rest of the thesis. It is not meant to be an exhaustive summary of statistics, machine learning, graph theory and text analysis. The reader is directed to some of the many available textbooks in these areas if requiring a more thorough background. For the material on text analysis, natural language processing in particular, the online Natural Language Toolkit (NLTK) book provides an excellent starting point for implementing some of these tools in the easy-to-understand python programming language[4] [16]. For a deeper understanding of the particular techniques, the referenced material is more thorough. For the econometric material, the original publications are often good sources to gain an understanding of the methods. For the material on machine learning, [74] is an excellent book that covers a range of topics from a probabilistic perspective. The results from graph theory are easily understood from the original research papers.

The main technique adopted in this thesis is text analysis, most prominently

[4]The book is available for free online at the NLTK website, http://www.nltk.org/book

sentiment analysis - the task of inferring emotional connotation from text. The final experiment makes use of some natural language processing (NLP, a subset of text analysis) techniques in order to extract mentions of organisations in text. Several techniques not directly applied within the context of this thesis are also presented in this section, such as semantic models (Latent Semantic Analysis, Random Indexing) and topic models (Latent Dirichlet Allocation) so as to give the interested reader an idea of alternative methodologies that could have been made use of. The specific reasons for why a particular method is selected in preference of alternatives are presented as part of the relevant experiments.

Both the second and third experiments make use of *network analysis*, or *graph theory*, to study network properties. The graph techniques used can be considered fairly basic from the point of view of graphs, yet have proved successful when applied in combination with the other tools used in the experiments.

The time series indicators derived in all three experiments are fed into time series models to measure the degree of relationship with hypothesised dependent variables. In particular, all experiments make use of time series analysis and accompanying statistical tests of model specification and significance.

2.2.1 Text Analysis

Text analysis, intimately related to but often considered distinct from *text mining*, is an area of research using statistics, machine learning, computation linguistics and other methods to tackle business and research problems through analysis of unstructured text data. Common text analysis tasks include:

- *Information retrieval* - the task of finding relevant documents from a collection, or *corpus*, to a given information need [67].
- *Named entity recognition* - the task of identifying known entities within a document [16].
- *Sentiment analysis* - the task of identifying emotional connotation within a document. An extensive survey of the field is available in [85].

Common applications of text analysis can be found in a range of areas.

- Business intelligence - e.g., extracting intelligence about products and services, so called competitive analytics
- E-Discovery - searching litigation documents for relevant information
- National security - monitor online data sources for national security risks

- Scientific discovery - derive new scientific patterns, especially in life sciences
- Sentiment analysis for finance, economics and business - e.g., using sentiment to forecast stock markets
- Social media monitoring - monitor brands and services on social media platforms

The primary focus of this thesis can be considered part of sentiment analysis, as it is primarily an attempt to extract emotional signals from text databases.

Sentiment Analysis

The study of sentiment analysis is a relatively young area of research that tries to apply natural language processing, computational linguistics and machine learning to determine the sentiment expressed in a given text [85]. This task is often simplified to a classification problem; that of classifying a given text by its sentiment polarity, positive or negative, or on a more granular level (e.g., one to five 'stars'). Early work in this area [83], [116] focused on product and movie review classifications. Pang's work indicates that such sentiment classification tasks do achieve lower accuracy than traditional topic classification using the same classification algorithms, showing the inherent difficulty in computationally determining sentiment polarity. Different approaches to the classification problem are discussed, such as comparing the accuracy of human constructed word dictionaries with features extracted from already labelled documents, showing that automatic feature extraction tend to outperform manual wordlist constructions on a given dataset. The best classification accuracy reported in their study was 82.9%, in a case where the algorithm looked for the presence of individual words, unigrams, and implemented classification using a support vector machine. Interestingly, the simple method of extracting 7 positive and 7 negative word indicators using a combination of human inspection and word frequency analysis achieved a classification accuracy of 69%. Further work in the area has focused on a finer sentiment classification, such as labelling a document on a sentiment scale [82], [101] and more accurate targeting of the sentiment and the referenced object [17]. Pang and Lee improved on their best sentiment classification performance by adopting a two-stage process where they first identified the sentences within a document likely to have subjective (emotional) content, to classify the documents by their sentiment polarity after excluding the remaining 'objective' sentences [84]. Their classification accuracy increased from 82.8% to 86.4%.

Another branch of work focuses on the prediction of market variables using sentiment features extracted from texts. Such studies often attempt to quantify sentiment using text-based data sources such as corporate reports and news media

with a view to predict stock prices, e.g. [107], [108], house prices [102], and corporate earnings [106], [65]. In [107] and [106] the authors make use of the General Inquirer system to process the text documents and the various Harvard dictionaries of emotion words. In [65] the authors create their own word dictionaries after showing that the Harvard lists aren't ideal for analysing financial texts. In [102] the author creates her own lists from the Harvard dictionaries for the same reason. In summary, there are a mix of previous studies employing machine learning techniques for classification and prediction and wordlist frequency techniques.

All the above studies had the luxury of pre-labelled documents available to them or a specific variable to predict. This is a critical observation. With pre-labelled documents it makes intuitive sense (disregarding the issue of whether or not such mined predictions carry over to unseen data) to look for the most accurate classification algorithm for the target domain (even if that involves mining for the features). However, when such labels are not available, automatic feature extraction is not only hard to implement (a proxy for the labels must be used, or if some labelled data is available it might be possible to use semi-supervised machine learning techniques to infer some of the labels) but is also difficult to validate. It is also well known that the accuracy of such classification algorithms tend to be very sensitive to the target domain.

Natural Language Processing

Natural Language Processing (NLP for short) is an area of research attempting to model language probabilistically to deal with the interaction between computers and natural languages, in order to extract the meaning from text or to generate text. As such, it can be considered an area of research in human-computer-interaction.

Tokenisation Before much analysis of text, it must first be split into smaller chunks, pieces or *tokens*. Often the tokens are individual words or sentences. These tasks are often more complicated than what might be expected because of punctuation (abbreviations that might contain punctuation), etc. The tokens can then be grouped into so called N-grams for further modelling. N-grams are sequences of N consecutive tokens, making up the building blocks of many probabilistic models in computational linguistics. When $N = 1$, they are called unigrams, when $N = 2$ they are called bigrams and when $N = 3$ they are called trigrams. N-grams are also used in other fields of computer science, such as data compression.

The Vector Space Model (VSM) The vector space model was developed so that documents can be analysed mathematically. Most semantic models, such as

Latent Semantic Analysis, can be considered extensions of the more basic vector space model. The vector space model was first used in the SMART (System for the Mechanical Analysis and Retrieval of Text) Information Retrieval System developed at Cornell University in the 1960s [93]. The VSM models documents as vectors in N dimensional space, where N is the total number of words appearing in the full corpus. Given a vocabulary of N words, a particular document can be modelled by a vector for which each entry corresponds to a score of a given word. There are several ways to assign such a score. The entry could be given by the total number of times the word appears in the document, tf (for 'term frequency'), the logarithm of tf (to reduce impact of highly frequent words), a one or a zero entry depending on the presence or not of the word in the document, or more commonly a score that reflects both the frequency of the word in the document and its ability to distinguish between, or 'separate', the documents in the full corpus $tf \times idf$ (for 'term frequency $\times$ inverse document frequency'). The inverse document frequency is commonly computed by taking the logarithm of the ratio $\frac{D}{df}$, where D is the total number of documents in the corpus and *df* (for 'document frequency') is the total number of documents the word appears in. There are several variations on both the tf and *idf* scores.

Given vector representations of the documents, the similarity between any pair of documents can be computed using linear algebra techniques. A common measure of similarity is given by the inverse of the angle between the two vectors in N dimensional space. If the document vectors are normalised to unit length, the cosine of the angle between the two vectors is given by the dot product of the two vectors, giving a similarity score between the vectors ranging from 0 to 1.

Latent Semantic Analysis (LSA) One major limitation of the simple vector space model when used in an information retrieval task, i.e., where the objective is to find a 'smaller' set of documents from a large collection that are somehow related to a specific user need represented as a short free text query, is the concept of *synonymy*. It is commonly found that multiple words can be used to refer to the same underlying object, implying that for any given query there will be a large number of potentially relevant documents that cannot be compared to the given query. Deerwester et. al. [29] applied a technique known as *singular value decomposition* (SVD), intimately related to *principal component analysis* (PCA, which is covered in a later section), to reduce the dimensionality of the word-by-document frequency-matrix M, where the entry $m_{i,j}$ is given by the frequency of word i in document j. It turns out that the lower-dimensional representation of word vectors better models the relationship between synonyms than the original rows of the matrix M. Two synonyms tend to have similar latent factors because they tend to co-occur with other similar words.

Similarly, lower dimensional vectors can be used to represent the documents in the collection.

Singular value decomposition of the matrix M is a process that decomposes the matrix into a product of three other matrices $M = U\Sigma V^T$, such that U and V have orthonormal columns (i.e., each column is orthogonal to all other columns, have inner product equal to 0, and have unit length), and Σ is a diagonal matrix. Matrices U and V are known as the *left* and *right singular vectors* of M respectively, and Σ contains the *singular values*. This decomposition can be shown to be unique up to reordering of the singular values. By convention, the matrix Σ is ordered such that the largest singular values appear from left to right. The interesting property of this decomposition is that it allows a rank k approximation of the original matrix M by simply replacing all but the k largest singular values in Σ by zero entries and carry out the sequence of matrix multiplications. The new product $\hat{M}$ can be shown to be the nearest, in the least square sense (i.e., the sum of squared errors is minimised), rank k approximation of the original matrix M. A corresponding composition of $\hat{M}$, $\hat{M} = \hat{U}\hat{\Sigma}\hat{V}^T$, can of course be made simply by deleting the relevant columns of U and V corresponding to the zero entries in the reduced $\hat{\Sigma}$. In this lower dimensional representation, the vectors of $\hat{U}$ and $\hat{V}$ correspond to word vectors and document vectors respectively.

Random Indexing Random projection, also known as random indexing [90], is a dimensionality reduction technique based on projecting the word-by-document frequency matrix onto a random subspace of much smaller dimension. In particular, given the $N \times D$ matrix M, where the entry $m_{i,j}$ is given by the frequency of word i in document j, a lower dimensional representation, $\hat{M}$, can be obtained simply by multiplying the matrix by a random $D \times D^{'}$ matrix T, where $D^{'} < D$. The columns in T should contain a small number of non-zero entries set to $+1$ or -1. There are several benefits to random indexing over traditional LSA.

- Computational complexity is much lower for RI
- The model can be incrementally updated without the need to process the full dataset
- RI does not require that the full matrix M is kept in memory but can simply process the documents one at a time in an online fashion

To clarify the third point, the matrix operation $M \times T$ can be processed online as,

$$w_i = \sum_{d_j | w_i \in d_j} rand(d_j) \tag{2.1}$$

where w_i is a given word in the vocabulary and d_j is a document in the corpus. The function $rand(d_j)$ simply allocates a random vector of dimension D'. The equation simply states that the word vector is represented as the sum of random document vectors for each document it appears in. This is an extremely efficient ($O(D)$) method to model the word by document co-occurrence matrix implicitly in reduced form (as one chooses D' before the algorithm starts). However, the resulting vectors cannot model synonyms. Random indexing, in the form above, cannot model associative similarity. Imagine a set of three words, X, Y, Z, for which X co-occurs with Y and Y co-occurs with Z but X never occurs in the same document as Z. RI cannot infer that X and Z are also somewhat similar as they share no terms in 2.1. This is a significant problem as it means RI cannot model synonyms, words never frequently occurring in the same documents but with identical meaning. Other topic models such as LSA and LDA (see the following section) do not suffer the same problem as latent (as opposed to random) factors are derived from the co-occurrence matrix.

Latent Dirichlet Allocation (LDA) *Latent dirichlet allocation* (LDA) is a generative probabilistic model of a corpus of documents [18]. LDA is part of a family of models known as *topic models*. LDA tries to overcome potential limitations of LSA and related methods that are derived from LSA. Most notably LDA tries to overcome the limitation of not being able to model polysemy (the property that a word might have several distinct meanings), as well as the lack of intuitive representations of the latent factors. In LDA, each document is modelled as a mixture of a finite number of possible topics. Each topic is in turn characterised by a distribution over words. The model assumes the following generative process for each document d in the corpus (i.e., follow this procedure to generate a random document under an LDA model),

1. Choose $N \sim \text{Poisson}(\phi)$
2. Choose $\theta \sim \text{Dir}(\alpha)$
3. To generate all the N words w_n:
   (a) Choose a topic $z_n \sim \text{Multinomial}(\theta)$
   (b) Choose a word w_n from $p(w_n|z_n, \beta)$, a multinomial probability conditioned on the topic z_n

The Dirichlet distribution $\text{Dir}(\alpha)$ has fixed dimensionality k, meaning the number of topics is predetermined. The word probabilities are parameterised by a $k \times V$ matrix β such that $\beta_{j,i} = p(w_i = 1|z_j = 1)$, given a total vocabulary size V. In other words, the LDA model describes the document collection as a probability distribution of words for each of the k topics, and a distribution of topics for each of the documents.

Quantitative Narrative Analysis Quantitative narrative analysis (or quantitative text analysis) is a fairly new branch of text analysis that seeks to model relationships between entities in text documents as networks, often extracting subject-verb-object triplets using natural language processing tools. Such semantic networks can then be studied using graph theory. Examples of such semantic networks in the literature can be found in [33] and [104]. A similar technique will be used in Chapter 7 to model the evolution of the structure of narratives in news.

Part-Of-Speech Tagging *Part-of-speech tagging* is the process of annotating words by their grammatical use. These algorithms often use dictionary lookup as well as probabilistic methods to annotate words by their part of speech. Stochastic methods are required since most words can have different part-of-speech tags depending on their context.

Named Entity Extraction *Named entity extraction* is the task of assigning entity type labels to tokens found in a document. For example, to infer if a token is a person, an organisation, a location, a date, or other entity. This is often achieved through a combination of dictionary lookup tables and rule-based annotation.

Co-Reference Resolution *Co-reference resolution* is the task of determining when several entities mentioned in a document really refer to the same entity, including inferring when a pronoun refers to a previously named entity.

GATE's ANNIE Process This section will clarify the methods used in Chapter 7 to perform sentence tokenisation and extract named entities (in particular, mentions of organisations). ANNIE - *A Nearly-New Information Extraction System* is the standard NLP processing 'pipeline' present in the Java GATE (*General Architecture for Text Engineering*) text analysis software.[5] ANNIE was originally developed by Hamish Cunningham, Valentin Tablan, Diana Maynard, Kalina Bontcheva, Marin Dimitrov and others. The sequence of steps taken by ANNIE to produce sentence and named entity annotations on a document can be summarised as,

1. Tokenise text
2. Entity detection based on dictionary lookup
3. Sentence splitter

[5]The full documentation of GATE can be found here https://gate.ac.uk/sale/tao/tao.pdf and documentation of ANNIE in Chapter 6 of the documentation, pp. 119-137.

4. Part-of-speech tagger

5. Entity detection using rules based on earlier annotations

6. Coreference

The text tokeniser is rule-based and uses pattern matching to decide how to annotate a token. The module can identify several different types of tokens - words, numbers, symbols, punctuation and space tokens.

Step 2 annotates tokens based on lists of known entities.

Step 3 relies on compositions of finite-state transducers that maps input text into a sequence of sentences. The splitter makes use of dictionaries of common abbreviations when deciding to split or not. The dictionaries ensures that common abbreviations such as U.S.A. do not lead to unwanted sentence splits.

In step 4, the part-of-speech tagger used by ANNIE [52] is a modification of the more well-known Brill tagger [23]. The tagger is dictionary and rule-based, as opposed to alternative taggers based on, e.g., Hidden Markov Models. Both the dictionary and the rules have been learned from a large corpus from the Wall Street Journal and is therefore suitable for analysis on news data (as in Chapter 7). The tagger proceeds in several steps. An initial processing step assigns tags to each word based on its most likely tag without consideration of context. Further steps correct these initial tags by applying trained rules which might for example change the tag of a token if it is preceded by some token of a specific tag. The trained model is compact in the sense of having only a few hundred rules and is therefore robust to over-fitting on the training data [52].

Step 5 relies on rules applied to the annotations produced in the earlier steps. Different types of entities can be identified, e.g., Person, Location, Organisation, Money, Percent, Date and Address entities.

Step 6 is perhaps the most involved part of the ANNIE processing chain. The task of this step is to match named entities from the previous step whenever they are believed to refer to the same entities.

2.2.2 Graph Theory

Graph theory is a branch of mathematics that deals with operations on graphs. A graph G is defined as a set of vertices V and a set of edges $E = \{(v_i, v_j) | v_i, v_j \in V\}$, where (v_i, v_j) denotes an edge from vertex v_i to vertex v_j for some $i, j \in [1, N]$ for $N = |V|$. If $(v_i, v_j) = (v_j, v_i)$ the graph is said to be undirected, otherwise the graph is said to be directed. For all the analysis in this thesis the graphs will be treated as undirected, henceforth this will therefore be assumed. The graph can be described by

a matrix A, for which $A_{i,j}$ is either one or zero indicating whether or not $(w_i, w_j) \in E$ respectively. The matrix A is called the *adjacency matrix* of G. This section will summarise the graph measures and algorithms used in later chapters.

Graph Density

Density is a measure of the proportion of edges to vertices of a given graph. It is defined by Equation 2.2 and is an increasing function of the total number of edges $|E|$ and decreasing in the total number of vertices $|V|$.

$$\frac{2|E|}{(|V|-1)|V|} \tag{2.2}$$

Global Clustering Coefficient or Transitivity

The *global clustering coefficient* of a graph (also commonly known as *transitivity*) is a measure of how strongly clusters are formed in the graph [75]. This is measured by the degree of vertex transitivity. Transitivity measures the probability that $(v_i, v_j) \in E$ given that $(v_i, v_k) \in E$ and $(v_j, v_k) \in E$ for some k. There are some slight variations on this definition, for example the local clustering coefficient. However, this thesis will only make use of the global coefficient defined as,

$$\text{transitivity} = \frac{3 \times \text{ number of triangles in the network}}{\text{number of connected triples of vertices}} \tag{2.3}$$

The number 3 accounts for the fact that there are 3 connected triples for each triangle. Here, a triangle is a fully connected triple of vertices and a connected triple is a set of three vertices with at least 2 edges between them. Intuitively, transitivity measures the fraction of triples that also have their third edge filled in to complete a triangle.

Edge Betweenness

The *betweenness* score of a given edge $(w_i, w_j) \in E$ of a graph G is defined as the number of shortest paths between any two pairs of vertices that run through it [43]. A shortest path is simply defined as the shortest sequence of edges running from one vertex to another. If there are more than one shortest path between two vertices they are typically all given equal weight so that the sum of weights equals one.

Graph Clustering

There exist several algorithms to cluster graphs into groups of nodes. This thesis has made use of two common algorithms. Part of the analysis in Chapter 6 will make use

of an algorithm by Newman [76] and Chapter 7 will make use of an algorithm by Girvan and Newman [43].

The algorithm in [76] aims to minimise the objective function given by,

$$Q(s) = \frac{1}{2m}\sum[A_{i,j} - P_{i,j}]\delta_{w_i,w_j} = \frac{1}{4m}s^T Bs \propto s^T Bs \tag{2.4}$$

where A is the adjacency matrix, P is a matrix defining a model of the edge distribution, $B := A - P$, m is the total number of edges in the network and

$$\delta_{w_i,w_j} = \begin{cases} 1, & \text{if vertex } w_i \text{ and } w_j \text{ are in the same cluster} \\ 0, & \text{otherwise} \end{cases} \tag{2.5}$$

describes the clustering to be found. The function δ_{w_i,w_j} may be written as $\frac{1}{2}(1 + s_i s_j)$ for some $s \in \{\pm 1\}^n$, where n is the number of vertices, when restricting the clustering identification to two groups. The problem becomes to optimise over all vectors s. This problem is NP hard so that an approximated linear algebra solution is derived. To generalise over more than two groups an iterative process is established.

The algorithm in [43] makes use of the concept of edge betweenness to determine community boundaries. The intuition behind the algorithm is that edges between communities are likely to have very high edge betweenness scores since most shortest paths between vertices in different communities tend to go via these edges. The algorithm yields a hierarchical clustering and proceeds as follows [43]:

1. Calculate the betweenness for all edges in the network.
2. Remove the edge with the highest betweenness.
3. Recalculate betweenness scores for all edges affected by the removal.
4. Repeat from step 2 until no edges remain or a given number of edges have been removed.

Social Network Analysis

Social Network analysis refers to the study of network interactions in social systems, which can now be considered a key technique in modern sociology. These social systems consist of agents (e.g., individuals or institutions) which are naturally connected to each other in some way, forming graphs. Examples of commonly studied systems include networks constructed from social media data, such as Facebook and Twitter [105] but also link relations of various types, such as on the World Wide Web and academic citations. Frequently used metrics of social network graphs include, but

are not limited to, connectivity metrics, distributional properties and clustering properties.

2.2.3 Statistics and Machine Learning Tools

This section presents the statistical tools primarily used in Chapters 4, 5, 6 and 7.

Mean, Standard Deviation and Coefficient of Variation

Given a probability distribution P, there are a number of useful summary statistics describing the distribution. Only sample empirical distributions are considered in this thesis, so the definitions will here be restricted to the various unbiased sample estimates of the true statistics.

The *sample mean* of an empirical distribution $P = \{x_1, \ldots, x_n\}$ of $n \in \mathbb{N}$ sample values, also known as the *expected value* $E[P]$, is defined as,

$$\hat{\mu} = \frac{\sum_{i=1}^{i=n} x_i}{n} \tag{2.6}$$

The *sample standard deviation*, $\hat{\sigma}$, is defined by,

$$\hat{\sigma^2} = \frac{\sum_{i=1}^{i=n} (x_i - \mu)^2}{n-1} \tag{2.7}$$

With these two definitions it is possible to define a standardised measure of the variability of a sample distribution, called the *coefficient of variation*,

$$\hat{CV} = \frac{\hat{\sigma}}{\hat{\mu}} \tag{2.8}$$

The coefficient of variation is a useful tool to compare the variability of two different sample distributions.

Bootstrap & Monte Carlo

Any statistic observed in a finite sample is subject to estimation and measurement errors. For example, it might not be possible to accurately measure the variables of interest or the observations at hand are more or less representative of the set of all possible observations. To account for such potential errors and estimate the confidence or significance of a particular statistic computed on the sample of observations at hand, it is possible to apply non-parametric techniques such as *bootstrap* [41] and *Monte Carlo* [73].

Bootstrap *Bootstrap* methods rely on random sampling with replacement from the set of observations to generate a sampling distribution of a statistic of interest. Given a set of observations from some random variable X, $X_i = x_i$ for $i = 1, \ldots, n$, let $R(x_1, \ldots, x_n)$ be a statistic of interest (e.g., the sample mean). Then the bootstrap procedure is used to generate a distribution $\hat{R}(\hat{x_1}, \ldots, \hat{x_n})$ by sampling $\hat{x_i}$ from the set $X_i = x_i$ for $i = 1, \ldots, n$. The distribution can be used to estimate quantities of interest such as the mean and variance of the statistic, $E[\hat{R}(\hat{x_1}, \ldots, \hat{x_n})]$, $Var(\hat{R}(\hat{x_1}, \ldots, \hat{x_n}))$.

Monte Carlo *Monte Carlo* methods are similar to bootstrap in that they rely on randomisation to estimate various numerical results. Monte Carlo simulations typically sample from a probability distribution to produce a large number of potential outcomes of a given variable. It is distinct from bootstrap in that bootstrap is typically defined as a *resampling* technique. In the context of this thesis, Monte Carlo simulations will be used to estimate empirical distributions of text analytic measures by repeatedly sampling without replacement from a large database of text documents.

Stochastic Processes, Stationarity and Order of Integration

A time series is an ordered sequence of data points. A stochastic process can be represented as a time series for which each datapoint is a observation from a random variable. As such, a stochastic process is a collection of random variables and an observed sequence is a time series. A stochastic process might be path dependent, i.e., a random variable might be dependent on some previous value.

A stationary stochastic process is a process for which the joint probability distribution does not change when it is shifted in time. This implies, in particular, that the mean and variance of the process must be constant over time. Therefore a stationary process cannot follow any trends. The concept of stationarity is very important when testing for a statistical relationship between two different stochastic processes.

Formally, a stochastic process $\{X_t\}$ is said to be stationary if any of its joint probability distributions have the same cumulative distribution function. In other words, let $F(x_{t_1+\tau}, \ldots, x_{t_k+\tau})$ be a cumulative distribution function of $\{X_t\}$ for some $t_1 + \tau, \ldots, t_k + \tau$, some k and some τ. Then $\{X_t\}$ is said to be stationary if for all k, for all τ and for all $t_1, \ldots, t_k$,

$$F(x_{t_1+\tau}, \ldots, x_{t_k+\tau}) = F(x_{t_1}, \ldots, x_{t_k}).$$

A stochastic process is said to be *covariance stationary* if $E[X_t] = \mu$ for all t (i.e., all random variables have the same mean), and $cov(X_t, X_{t+k}) = \gamma_k$ for all t and k (i.e.,

the k'th lag auto-covariance is independent of t and depends only on the order of the lag). Thus, a stationary stochastic process must be covariance stationary, but the converse is not always true. In both definitions of stationarity, a series is said to be trend-stationary if removing a deterministic trend from the data turns it into a stationary series.

A stochastic process is said to have *order of integration* d (or to be integrated of order d) for some $d \in \mathbb{N}$ if d successive first order difference operations are needed to arrive at a covariance stationary series. A series which is integrated of a higher order than zero is said to have a unit root. There are several tests for the existence of unit roots. Stochastic processes with such roots are hard to model accurately in regressions, see the section on spurious regressions for more details of the problem.

Augmented Dickey-Fuller Test The Augmented Dickey-Fuller test [91] is a popular test for the existence of a unit root. The test is applied to the following model,

$$\Delta y_t = \alpha + \beta t + \gamma y_{t-1} + \delta_1 \Delta y_{t-1} + \cdots + \delta_{p-1} \Delta y_{t-p+1} + \epsilon_t,$$

where α is a constant, β represents a time trend, and p is the lag order of the process. The lag order p needs to be determined before the test is carried out.

The unit root test is then carried out under the null hypothesis $\gamma = 0$ against the alternative hypothesis of $\gamma < 0$. If the null hypothesis cannot be rejected this implies the existence of a unit root.

Kwiatkowski-Phillips-Schmidt-Shin Test The Kwiatkowski-Phillips-Schmidt-Shin (KPSS) [59] test is a test for the null hypothesis that a stochastic process is stationary around a deterministic trend. The test is applied to a model which is the sum of a deterministic trend, a random walk process and a stationary error term. The null hypothesis is that the variance of the random walk term is zero, which corresponds to trend stationarity. Rejecting the null therefore implies that the original series is non trend-stationary. In other words, the test is applied to a model of the following form,

$$y_t = \beta t + r_t + \epsilon_t,$$

for a constant β representing the deterministic trend, ϵ_t is a stationary error term, $r_t = r_{t-1} + u_t$ is a random walk process with u_t i.i.d with zero mean and constant standard deviation σ_u^2. r_0 is treated as a fixed constant and therefore serves the role of an intercept [59]. Since e_t is assumed stationary, the hypothesis that $\sigma_u^2 = 0$ is equivalent to the hypothesis that y_t is trend-stationary. If β is fixed as zero, the null

corresponds to the null hypothesis that y_t is level stationary (around r_0) as opposed to stationary around the trend.

Pearson Correlation

The *Pearson correlation coefficient*, often referred to as Pearson's r, is the most widely used statistical measure of linear dependence between two random variables X and Y [74]. The coefficient can take values from -1, for a completely negative dependence, to +1, for a completely positive dependence. Let $E[X] = \mu_x$, $E[Y] = \mu_y$, $\sqrt{E[(X - \mu_x)^2]} = \sigma_x$ and $\sqrt{E[(Y - \mu_y)^2]} = \sigma_y$ then the Pearson correlation coefficient is defined by,

$$\frac{E[(X - \mu_x)(Y - \mu_y)]}{\sigma_x \sigma_y} \tag{2.9}$$

Spearman Rank Correlation

The *Spearman's rank correlation coefficient* , often referred to as Spearman's rho, is a measure of dependence between two ranked variables X and Y. The coefficient can take values from -1, for a completely negative dependence, to +1, for a completely positive dependence. This measure of correlation is therefore nonparametric as opposed to Pearson's r which is a parametric test. Let x_i be the ordered rank of observation X_i and y_i be the ordered rank of observation Y_i, then the Spearman rank correlation coefficient is defined by,

$$1 - \frac{6 \sum d_i^2}{n(n^2 - 1)} \tag{2.10}$$

where n is the number of observations and $d_i = x_i - y_i$.

Linear Regressions

One of the most widely used regression models is *linear regression*. The model assumes that the output variable Y (also known as the dependent variable) is a linear function of the $N \in \mathbb{N}$ input variables $X^1, \ldots, X^N$ (also known as independent variables or explanatory variables). The model is written as follows,

$$y_i = \alpha + \beta_1 x_i^1 + \cdots + \beta_n x_i^n + \epsilon_i \tag{2.11}$$

for some constant α, input weights β_i for $i = 1, \ldots, N$ and error term ϵ, most often assumed to be i.i.d. Gaussian [74] with constant variance σ^2 and zero mean. The assumption of constant error variance is known as homoscedasticity, and the absence

as heteroscedasticity. P-values on the significance of the constant term and the input weights are usually not defined if this assumption does not hold in practice. The equation 2.11 is most commonly estimated using *ordinary least squares* (OLS, see the next section). Once α and β_i, for $i = 1, \ldots, N$ have been estimated, predictions of the response variables y_i can be generated from the fitted equation $\hat{y}_i = \alpha + \beta_1 x_i^1 + \cdots + \beta_n x_i^N$. The errors of the predictions $\hat{u}_i = y_i - \hat{y}_i$ are known as *residuals*.

How well the explanatory variables can be combined linearly to predict the output variable Y, is commonly measured by the *coefficient of determination*, R^2,

$$R^2 = 1 - \frac{SSR}{SST} = \frac{SSE}{SST}, \tag{2.12}$$

where $SST = \sum_{i=1}^{T}(y_i - \bar{y})^2$ is the *total sum of squares* (here $\bar{y} = \frac{1}{T}\sum_{i=1}^{T} y_i$ represents the average of y_i), $SSE = \sum_{i=1}^{T}(\hat{y}_i - \bar{y}_i)^2$ is the *explained sum of squares* by the predictions of the equation, $\hat{y}_i$, and $SSR = \sum_{i=1}^{T} \hat{u}^2$ is the residual sum of squares. The coefficient of determination therefore measures the proportion of total variability in the response variable that is accounted for by the equation. Notice that this number can be made arbitrarily large simply by the inclusion of more explanatory variables X_j. The *adjusted* R^2 adjusts this coefficient by a number that penalises for the number of free variables, N, in the equation,

$$adj.R^2 = 1 - \frac{SSR}{SST}\frac{T-1}{T-N-1} \tag{2.13}$$

The significance of estimates of the coefficients can be tested under the null hypothesis $\beta_i = \bar{\beta}_i$ using a t-test. The t-statistic,

$$\frac{\hat{\beta}_i - \bar{\beta}_i}{se(\hat{\beta}_i)},$$

where $se(\hat{\beta}_i)$ is the standard error of the estimated coefficient, has a student t distribution with $T - N - 1$ degrees of freedom if the null hypothesis is true.

The standard error of a coefficient is most easily derived if equation 2.11 is rewritten in matrix form as,

$$y = \beta X + \epsilon, \tag{2.14}$$

where y is a $T \times 1$ vector representing the observations of the explanatory variable, β is a $(N+1) \times 1$ vector of coefficients where the first entry corresponds to α, and X is a $T \times (N+1)$ matrix of input variables where the first column is a vector where each entry equals 1. Under the assumption that the error terms are not heteroscedastic,

the variance of β can be estimated by,

$$\begin{aligned} Var(\hat{\beta}) &= \hat{\sigma}^2 (X^T X)^{-1}, \\ \hat{\sigma}^2 &= \frac{\hat{u}^T \hat{u}}{T - N - 1} \end{aligned} \tag{2.15}$$

which is derived from the OLS estimate, see the following section (equation 2.16), of the coefficients.

The standard error of an estimated coefficient is therefore given by the square root of the corresponding diagonal entry in $Var(\hat{\beta})$. This gives the t-statistic needed to estimate the significance of the null hypothesis for the value of any specific coefficient estimate.

Ordinary Least Squares (OLS) The equation 2.11 is most commonly estimated from the data using a method known as *ordinary least squares*. The objective function to minimise over the parameters α and β_i for $i = 1, \ldots, N$ is the residual sum of squares RSS. The ordinary least squares estimate of the coefficient vector in equation 2.14 can be shown to be,

$$\hat{\beta} = (X^T X)^{-1} X^T y \tag{2.16}$$

Spurious Regressions The seminal work by Granger published in *Econometrica* [45] outlines one of the main problems with interpreting regression results when assumptions of the model are violated. In particular, when the errors are autocorrelated. In that paper, they highlight three main problematic consequences of autocorrelated errors,

1. Estimates of coefficients are inefficient (in other words, there is greater variability in the OLS solution)
2. Forecasts based on equations are suboptimal
3. The usual significance tests on the coefficients of the model are invalid

The paper deals with the most serious of these three issues, the third one. The problem is framed in terms of equation 2.11. The null hypothesis is formulated as $\beta_1 = 0, \ldots, \beta_n = 0$, tested using test statistic,

$$F = \frac{T - (n+1)}{n} \frac{R^2}{1 - R^2},$$

where T is the number of observations of the variables. The significance of the F statistic is then compared to Fisher's F distribution. However, if such coefficients β can be found that satisfies 2.11, and Y itself is not a Gaussian (i.e., equal to the error term), the null hypothesis cannot be true (by assumption) and so to test it would be inappropriate. If on the other hand, Y is non-stationary and the null hypothesis is in fact true, then we must have $\epsilon_i = y_i - \alpha$, which is impossible since the right hand side will be non-stationary but ϵ_i is assumed to be Gaussian. The F statistic would not follow the assumed Fisher distribution. Granger and Newbold proceed by supplying Monte Carlo simulations for a range of common non-stationary series and show that the null hypothesis is rejected far too often if errors are autocorrelated [45]. The authors advice to always test the errors of an equation for autocorrelation using the Durbin-Watson test.

Durbin-Watson Test The *Durbin-Watson* test is a test for serial correlation of the error terms in a linear regression model [37], [38], [39]. The test statistic is given by,

$$d = \frac{\sum_{t=2}^{T}(\epsilon_t - \epsilon_{t-1})^2}{\sum_{t=1}^{T}\epsilon_t^2} \tag{2.17}$$

where T is the number of observations and ϵ is given as in 2.11. A value of d around 2 indicates no autocorrelation of the errors terms, a value less than 2 indicates presence of positive serial correlation, and if it is above 2 there is evidence of negative autocorrelation. Typically, a low value of d is evidence that the significance of the equation is potentially *overestimated*. A high value of d indicates the significance is potentially *underestimated*. However, the test is not applicable in those cases when lagged values of the dependent variable Y are included as explanatory variables. In this case, other tests can be used, such as the Breusch-Godfrey test.

Breusch-Godfrey Test The *Breusch-Godfrey* test [21] is a further test for the presence of higher order autocorrelation in the errors terms of a linear regression. The test is considered more powerful than the Durbin-Watson test and is applicable even for the case when lagged dependent variables are used as explanatory variables. The null hypothesis is that there is no autocorrelation in the errors terms up to a specified order p. Given the model in 2.11, let $\hat{u}_i$ denote the errors (residuals) of the fitted equation. The following equation is constructed,

$$\hat{u}_i = \alpha + \beta_1 x_i^1 + \cdots + \beta_n x_i^n + \rho_1 \hat{u}_{i-1} + \cdots + \rho_p \hat{u}_{i-p} + \epsilon_i \tag{2.18}$$

Define the null hypothesis of no serial correlation of the error terms as

$\rho_1 = \cdots = \rho_p = 0$. Breusch and Godfrey proved [21] that the fitted R^2 of this equation, scaled by $n = T - p$ (i.e., nR^2), where T is the number of observations, follows a Chi-Squared distribution with p degrees of freedom under the null. Thus, serial correlation of order up to p can be tested by comparing this statistic to the Chi-Squared distribution.

Breusch-Pagan Test The *Breush-Pagan* test [22] is a test for a linear form of heteroscedasticity of the error terms of a linear regression model. Given the linear regression equation 2.11 and the residuals $\hat{u}_i$, the test considers the model,

$$\hat{u}_i^2 = \alpha + \rho_1 x_i^1 + \cdots + \rho_n x_i^n + \epsilon_i \tag{2.19}$$

The dependence of the squared residuals on the values of the independent variables can be tested for given the null hypothesis $\rho_1 = \cdots = \rho_n = 0$. From this, a test statistic under the null hypothesis is derived which has a Chi-Squared distribution with n degrees of freedom [22].

Ramsey RESET Test The *Ramsey RESET* test [88] is a test for mis-specification of the functional form of a linear regression model. In particular, it aims to test if non-linear interactions of the independent variables should be included. The test proceeds by first fitting equation 2.11 to get estimates of the dependent variable $\hat{y}_t$. Powers of the estimated variable $\hat{y}_t$ are then added to a new equation,

$$y_i = \alpha + \beta_1 x_i^1 + \cdots + \beta_n x_i^n + \rho_1 \hat{y}^2 + \cdots + \rho_{p-1} \hat{y}^p + \epsilon_i \tag{2.20}$$

The null hypothesis is formulated as $\rho_1 = \cdots = \rho_{p-1} = 0$ and can be tested for using a standard F test. The intuition behind the test is that if the coefficient on some non-linear interaction term of the dependent variables is significantly different from zero, the original equation can be considered mis-specified. However, the exact source of the missing non-linearity will not be known.

Kolmogorov-Smirnov Test of Residuals Finally, the error terms of equation 2.11 are assumed to follow a Gaussian distribution. There are several tests for the equality of two distributions. One such nonparametric test is the *Kolmogorov-Smirnov* (K-S) test. The K-S test can be used to test for equality of two continuous sample distributions or one continuous sample distribution to a continuous reference distribution.

The test statistic considers the sample (given the samples $x_1, \ldots, x_n$) cumulative distribution function,

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}_{[-\infty,x]}(x_i),$$

where $\mathbb{I}_{[-\infty,x]}(x_i)$ is the indicator function equal to one if $x_i \leq x$ and zero otherwise. The sample distribution is compared to another sample distribution or a reference distribution $F(X)$ via $\sup_x |F_n(x) - F(x)|$. The form of the distribution of this test statistic has been derived and can be used to test the null hypothesis that the two distributions are the same [36].

Given this test for distributional equality, the normality assumption of the error terms of a linear regression can be tested for.

Vector Autoregression (VAR)

A *VAR* model is an extension of a particular type of linear regression model known as an autoregression (AR) model. In an autoregressive model of order p, AR(p), the response variable Y is represented as a sum of previous values and an error term, $y_i = \alpha + \beta_1 y_{i-1} + \cdots + \beta_p y_{i-p} + \epsilon_i$. A VAR($p$) model is a multidimensional autoregressive model where the dependent variable Y is a vector of random variables that all depend on each other inter-temporally. Each component in Y therefore has its own equation describing it in terms of p lags of itself and the other components in Y. These models will not be used extensively in this thesis, except to perform Granger-causality tests, in the two dimensional case.

Multivariate Portmanteau Test The Portmanteau test statistic is used to test for the absence of serial correlation in the error terms of a VAR(p) model. The statistic used is defined as,

$$\begin{aligned} T\sum_{j=1}^{h} tr(C_j^T C_0^{-1} C_j C_0^{-1}) \\ C_i = \frac{1}{T}\sum_{t=i+1}^{T} \hat{u_t} \hat{u_{t-i}}^T \end{aligned} \tag{2.21}$$

where $\hat{u_t}$ is the error term of the VAR(p), h is the lag order considered in the test, and T is the total number of observations. The test statistic is approximately distributed as Chi-Squared with $K^2(h-p)$ degrees of freedom under the null hypothesis of no serial correlation of the residuals, where K is the dimension of the VAR model.

Multivariate Breusch-Godfrey Test The multivariate Breusch-Godfrey test is similar to the univariate test. It is based on the following equation,

$$\hat{u_i} = A_1 y_{i-1} + \cdots + A_l y_{i-l} + \rho_1 \hat{u}_{i-1} + \cdots + \rho_h \hat{u}_{i-h} + \epsilon_i, \tag{2.22}$$

where l is the lag order of the VAR and h is the order of serial correlation to be tested for. The null hypothesis is given as $\rho_1 = \cdots = \rho_p = 0$. The restricted version of equation 2.22 is defined as the equation for which $\rho_1, \ldots, \rho_p$ are forced to be zero. Whereas the original version is considered an unrestricted version. Based on those definitions, one common test statistic used for the Breusch-Godfrey test is given by,

$$T(K - tr(\hat{\Sigma}_R^{-1}\hat{\Sigma}_e)), \tag{2.23}$$

where $\hat{\Sigma}_R^{-1}$ and $\hat{\Sigma}_e$ are the covariance matrices of the residuals from the restricted and unrestricted equations respectively. This statistic is approximately distributed as Chi-Squared with hK^2 degrees of freedom, where K is the dimension of the VAR model.

Wald Test Another common test of parameter restrictions on a linear regression equation, like the F-test, is the *Wald* test [49]. In the multivariate case, when restrictions on several parameters of equation 2.14 are to be tested, the null hypothesis can be formulated as $R\beta = r$ for the parameter vector β, a $Q \times (N+1)$ matrix R describing Q potential restrictions on the $N+1$ parameters. The test statistic is given by,

$$(R\hat{\theta}_n - r)^T [R(\hat{V}_n/N)R^T]^{-1}(R\hat{\theta}_n - r), \tag{2.24}$$

where $\hat{V}_n$ is the covariance matrix of X. The statistic approximately follows a Chi-Squared distribution with Q degrees of freedom.

Granger-Causality

The hypothesis that a given variable 'causes' another variable is encountered throughout the sciences. Does a given drug help in the treatment of a particular disease? Does intelligence make people better off financially? It is only possible to get truly satisfying answers to these questions if it is possible to intervene in the system and observe the consequences of that intervention. This is often not possible if

experiments cannot be reliably constructed and carried out. In particular it is often not possible when studying effects on the economy or in financial markets, unless a 'natural experiment' exists in which some intervention was made in the past and just happens to be available and present in the data. But even in this case, due to the high complexity of social systems, it is not always possible to explain changes in variables due to the said intervention.

Thus, the problem is often that of testing whether or not changes in one variable systematically predict changes in another variable (often without perfect knowledge of the system producing the two variables), beyond what can be expected because of noise and sample size. A common method pioneered by C. Granger and R. F. Engle, known as *Granger causality*, is based on this fundamental observation [44]. In its most basic form, Granger causality tests for a causal relationship from random variable X to variable Y by comparing a linear regression of Y using only lagged dependent values to a regression of Y using both lagged dependent values and lagged values of variable X. Testing whether all the coefficients on lagged values of X equal zero is the most basic form of the test. In other words, the following two equations are compared.

$$\begin{aligned} y_i &= \alpha + \beta_1 y_{i-1} + \cdots + \beta_p y_{i-p} \\ y_i &= \alpha + \beta_1 y_{i-1} + \cdots + \beta_p y_{i-p} + \gamma_1 x_{i-1} + \cdots + \gamma_p x_{i-p}, \end{aligned} \tag{2.25}$$

for some lag parameter p.

Spurious Granger-Causality Estimation of Granger-causality is made problematic in the presence of non-stationary time-series, as explained in [50]. Toda and Yamamoto [109] showed that it is only possible to rely on test statistics if 'extra' lags (where the number of extra lags to include is given by the maximum order of integration of the two series) of the two variables are thrown into the equation before a Wald test is carried out on the original lags of the model. They showed that adding such extra lags ensures that the test statistic is normally distributed and inferences can therefore be made as usual (i.e., when both series are stationary).

The T-Y approach for Granger-causality testing will be followed to avoid such problems. The main steps of the procedure are as follows:

1. Determine the order of integration of the two series involved. Apply the Augmented Dickey-Fuller (ADF) test for the existence of a unit root and the Kwiatkowski-Phillips-Schmidt-Shin (KPSS) test for the null hypothesis that a series is level or trend stationary. Repeatedly difference the series until both tests suggest stationarity of both series. Let m denote the highest order of

integration of any of the two series found (i.e., the maximum number of successive difference operations needed for the series to become stationary).

2. Find the number of lags p to include in the VAR model of the two variables. The Akaike Information Criterion (AIC) [13] is used to select p.

3. Check for parameter stability of the VAR by inspecting OLS-CUSUM plots [87].

4. Test the VAR model for potential miss-specification, in the form of serially correlated residuals, by applying the multivariate Portmanteau- and Breusch-Godfrey tests. If either of the two tests reject the null with a p-value of less than or equal to 5%, include further lags in the VAR until the residuals appear non-correlated. Let p now denote the total number of lags included after this step.

5. Perform the critical step suggested by Toda and Yamamoto. Add m extra lags to our $VAR(p)$ model. This ensures that the test statistic used follows the assumed distribution under the null.

6. Test for Granger non-causality in both directions between the variables by performing Wald tests. To test for Granger non-causality from variable X to variable Y the null hypothesis is that the coefficients on the first p lags of X in the regression with Y as the dependent variable are all equal to 0. This parameter restriction is tested for using a Wald test with p degrees of freedom. Finally, reject the null of Granger non-causality if the p-value is less than a specified thresholds (10%, 5% or 1%).

A Note on Correlation vs Causation

It is important to keep in mind that whatever method is used to infer causality between two existing time series without knowledge of the true data generating process or having the ability to interfere with that process, it is never possible to infer 'true causality' (or the truth about the existence of any relationship between two variables X and Y). The problem can be formulated mathematically in the form of the law of total expectations,

$$E_{Y|X}[Y|X] = E_Z[E_{Y|X}[Y|X, Z]|Z],$$

for any random variable Z. In other words, the expected value of Y given X (the prediction) is just the average of all such predictions made after factoring in any third variable Z. If X and Z are statistically dependent in some way, our prediction of Y

given X depends on our variable Z, and the possibility of the existence of such a dependent variable Z can never be excluded. Therefore predicting Y given variable X might work even if the two variables aren't related 'causally' in any way, but they could both be driven by some unknown third variable Z.

This makes it apparent that it is impossible to infer causality using inductive statistics alone, whatever method is used, but a theoretical argument must always support any such claim.

P-value Adjustments Whenever a number of hypotheses are tested at once there are several p-value adjustment strategies that can be used to account for the increased probability of a Type I error. For example, *Bonferroni correction* [9] is a method that attempts to control for the *familywise error rate* (FWE). The FWE is defined as the probability of making one or more false rejections in a "family" of hypothesis tests. Bonferroni correction of the significance level is performed by scaling it by the size of the "family". Bonferroni correction is considered a rather conservative strategy in that it greatly limits the number of Type I errors, but tend to introduce a lot of Type II errors. Thus, if a significance level of 0.05 is used, a test will be said to be significant if the p-value is less than $\frac{0.05}{m}$, where m is the total number of tests performed.

Another correction method, which is less conservative in nature, developed by Benjamini and Hochberg [15], aims at controlling for the *false discovery rate* (FDR). The FDR is defined as the proportion of falsely rejected null hypotheses within the total set of rejections.

Principal Component Analysis (PCA)

Principal component analysis (PCA) is an unsupervised machine learning technique to discover the main structure of a dataset. It is an orthogonal linear transformation that maps a data matrix of N items into a new coordinate system where the greatest variance is described by the first component, the second greatest variance by the second component, and so on. It is commonly employed as a dimensionality reduction technique by making use of only the first $K < N$ components of the new representation.

Mathematically, PCA can be described by the following theorem taken from [74] (where a proof is also provided),

Theorem 1. *Given the empirical data matrix X, the orthogonal set of L linear basis vectors $w_j \in \mathbb{R}^D$ that minimises the error function,*

$$\frac{1}{N}\sum_{i=1}^{N} ||x_i - \hat{x_i}||^2,$$

where $\hat{x_i} = Wz_i$ *for an orthonormal matrix of* loading coefficients, W, *and* score vectors, $z_i \in \mathbb{R}^L$, *are given by the* L *eigenvectors with the largest eigenvalues of the empirical covariance matrix,*

$$\frac{1}{N}\sum_{i=1}^{N} x_i x_i^T,$$

where it is assumed that the x_i *all have zero mean. Let* V_L *denote the matrix with these* L *eigenvectors as columns. The orthogonal projection of the data into the space spanned by the eigenvectors,* $V_L^T x_i$, *is the optimal lower dimensional embedding of the data.*

Furthermore, the direction of the eigenvectors of the largest eigenvalues are those for which the data show the highest variance.

When the observations in X are on substantially different scales, it is often common to standardise the variables before computing the covariance matrix, or equivalently to use the correlation matrix instead of the covariance matrix. It also makes more sense to use the correlation matrix when the relationship of interest between the original variables is linear correlation as opposed to covariance. In this case the principal component represents the greatest correlation as opposed to variance.

Information Theory

Information theory is an area of mathematics and computer science attempting to quantify and measure information. The field was pioneered and developed by American mathematician and engineer Claude E. Shannon, primarily in order to derive limits on data processing operations such as data compression. The field is very broad with applications in a range of scientific and practical areas. This section will only describe two ideas from the field, *Shannon entropy* and *Akaike information criterion*.

Shannon Entropy *Shannon entropy* is a measure of the predictability of a distribution [96]. Alternatively, it can be considered a measure of the degree of uncertainty or information present in a distribution. In other words, the extent to which a particular observation can be predicted ex ante. For a discrete probability distribution P it is defined as,

$$\mathrm{H}(X) = -\sum_i \mathrm{P}(x_i)\log_b \mathrm{P}(x_i), \tag{2.26}$$

where b is the base of the logarithm used, commonly chosen to be 2. The version of equation 2.26 for a continuous probability distribution replaces the summation across a finite number of observed values with an integral over all possible values,

$$h[f] = \mathrm{E}[-\ln(f(x))] = -\int_{\mathbb{X}} f(x)\ln(f(x))\,dx, \tag{2.27}$$

where $f(x)$ is the continues probability density function and $\mathbb{X}$ is the support of the function (i.e., the possible observations of the random variable X).

Akaike Information Criterion The *Akaike information criterion*, developed by Hirotugo Akaike in the early 1970s [13], is a criterion for model selection. It is not a means to determine the fitness of a particular model, but rather a means to compare how well different models fit observed data. The criterion relies on the *likelihood function* of a statistical model. The likelihood of a specific parameter value for a given model and observed data assumed to be generated from the model, is defined as the probability of observing a particular set of parameters, θ, given the observed data $X = x$,

$$\mathcal{L}(\theta|X = x) = P(X = x|\theta). \tag{2.28}$$

A common use of the likelihood function is to find the parameter θ^* that maximises the probability of the observed data, known as the maximum likelihood, L. This value is model specific and can sometimes be difficult to compute. In some cases, it is easier to compute the logarithm of the function 2.28 before taking derivatives in order to find the parameter that yields the maximum value.

Given θ^*, the AIC score for a given model is defined as,

$$\mathrm{AIC} = 2k - 2\ln(L), \tag{2.29}$$

where k is the number of parameters in the model. Thus, the AIC penalises models with more parameters and rewards models with greater maximum likelihood. When applying the criterion to multiple models, the model with the lowest AIC is most often chosen as the better model.

2.3 Discussion

This chapter presented relevant literature, divided into a theoretical review in the first section and technical review in the second section.

The theoretical review section presented the theoretical background material on

decision-making; starting with one of the most prominent models of decision-making, rational choice theory, and proceeding to present conviction narrative theory as an alternative theory of decision-making which acknowledges the role of uncertainty and serves to explain how agents are able to act in the face of it. After this section, a generic discussion about some of the potential problems with the statistical analysis of large data sources is presented. The discussion contrasts some of the differences between more traditional hypothesis testing and what is termed machine-driven hypothesis testing in the context of publication biases and independence of tests. In particular, it is argued that biased models tend to perform better on out-of-sample data (given the bias-variance tradeoff) and that theory-driven testing can be considered a biased model which remains fixed across multiple hypotheses, thus avoiding the potential issues when testing and evaluating hypotheses independently. The theoretical review section ends with a section that puts the text analytical methodology derived throughout the thesis in this context.

The technical review section presented the tools needed to develop the methodology and test the thesis hypothesis; text analysis, graph theory, statistics and machine learning.

3 Data

The ideal type of data source, for the purpose of identifying the emotional patterns that would be characteristic of CNT, is subjective, time-stamped, textual data, collected to avoid response biases, and stretching over a lengthy period, in which agents freely and naturally describe the actions they take and why. Ideally the source would comprise the views of multiple agents in contact with one another. It would then allow a computer-analytic method of constructing influence networks and developing sentiment patterns from the free text data. Unfortunately, although such data exist to a significant extent inside some organisations it has not been available for use within this thesis. Therefore, focus has been on some alternative data sources described in this chapter.

3.1 News

The most extensively used data source throughout the thesis is Reuters News archive. This was generously provided by Thomson Reuters at no cost to the Centre for the Study of Decision-Making Uncertainty at University College London. Reuters' historical news archive, consists of over 20 million news articles in several languages reaching as far back as 1996 [6].

In this thesis, the archive has been restricted to a much smaller subset consisting of those articles published by Reuters in New York or Washington. This archive of US

Reuters news contains a total of 2 219 647 articles from January 1996 through October 2014. These documents will be referred to as US RTRS.

3.2 Emails

Getting access to email databases is somewhat harder than getting access to news databases, primarily because of privacy issues. However, one of the case studies investigated in Chapter 6 makes use of the email archive of Enron Corp. This database was originally made public during the Federal Energy Regulatory Commission's (FERC) investigation into the energy trading market in May 2002, and is available in several forms and sizes. A MySQL version of the database is used, which mainly covers the period 2000-2002 after Enron's foremost rise but during its fall. This data were prepared by Adibi and Shetty [11], who have explained how the data have been modified and cleaned in several iterations. The data have been used in several research projects in different settings.

3.3 Analysts' Reports

Via collaboration with the Bank of England access has been arranged to a range of broker reports and market intelligence documents. This type of data is likely to be closer to the ideal datatype described in the introduction to this chapter as the content creators are likely less constrained on the degree of subjectivity they are allowed to express.

3.3.1 Broker Reports

An archive of 14 brokers from June 2010 through June 2013 consisting of documents of a primarily global economic focus will be analysed in Chapter 5. The archive consists of approximately 111 documents per month. The documents are very long (up to 50 pages in some cases), and therefore contain a large number of words. These documents will be referred to as BROKER.

3.3.2 Internal Bank of England Market Commentary

The Market Directorate of the Bank of England produces a range of internal reports on financial markets and the financial system, some of which provide high-frequency commentary on events and some of which provide deeper, or more thematic, analysis. In this thesis, some documents of the former kind, more specifically daily reports on the current state of markets are analysed in Chapter 5. These documents mainly

cover financial news and how markets appear to respond to such news. It is therefore expected that these documents correlate with financial sentiment. An average of 26 documents per month from January 2000 through July 2010 will be analysed. These documents will be referred to as MCDAILY ('Market Commentary Daily').

3.4 Economic and Market Data

During the course of the thesis a variety of economic and financial time series will be used. The US GDP data are taken from the Bureau of Economic Analysis http://bea.gov/. Financial data are retrieved from Yahoo finance [8]. Other economic data are taken from the Federal Reserve Bank of St. Louis database, FRED, https://research.stlouisfed.org/fred2/

4

Methodology and Preliminary Evaluation

Feelings of approach versus avoidance are key to understanding CNT. This chapter develops the RSS methodology used in the experiments. The chapter introduces and explores a number of methodological questions relating to the RSS measure. Several scientific contributions relating to the evaluation of word frequency sentiment analysis are presented in the *Preliminary Evaluation* section after the *Methodology* section.

The chapter presents the basic RSS methodology, in particular how the preselected lexicons were constructed, how they have been used to produce the RSS score and how articles are selected and aggregated into a time series of a given frequency. The second part of the chapter evaluates the methodology on several criteria. The evaluation section discusses the effects of different ways to scale the relative frequency counts of excitement and anxiety words, the potential effect of basic negation, as well as more non-trivial issues such as the temporal evolution of the relative ranks and frequencies of the words in a given list and, most importantly, the extent to which the words in a list tend to capture the same emotional signal over time. It is argued that the latter criterion can be used to evaluate the performance of word-frequency based sentiment analysis in general. There are other means by which an emotional lexicon can be evaluated. For example, a lexicon could be manually evaluated, at fairly low cost, using crowd-sourcing platforms such as Amazon's Mechanical Turk platform[1], as was done in [34].

[1] https://www.mturk.com/mturk/welcome

4.1 Methodology

Given the potential downsides of sentiment analysis based purely on machine learning techniques, as discussed in Chapter 2 in the section about sentiment analysis (in particular the necessity to have access to labelled training data), this study applies the method of manual, expert word list construction in order to create two dictionaries; one that is labeled 'excitement' and one labeled 'anxiety' that represent approach and avoidance emotions respectively.

The RSS word lists were created from the Harvard IV negative and positive words [2] and lists of negative and positive words generated by Loughran and McDonald [65], that were developed specifically for the analysis of financial texts. Within a research team comprising a social anthropologist, a sociologist and psychoanalyst and a clinical psychologist visual inspection suggested that these 'positive' and 'negative' lists were not exactly correlated with excitement (approach) and doubt and anxiety (avoidance) words. Therefore, these lists were edited by this research team [114] and some words added to produce more intuitive lists, exercising psychological judgment as to which words suggested attraction and which repulsion. The (emotional) interpretation of words is highly consistent across people and even across languages [118] and informed selections of word lists relating to emotional dimensions correspond well to independent judgments from naive raters [19], [60]. The method also has the advantage that it can be used across any type of text and has an intuitive interpretation. A lot of care was exercised to avoid including emotion words that could carry economic and financial interpretations, e.g., "recession", "bankruptcy" and "crisis", which tend to exhibit high frequencies during specific periods - potentially introducing problems of *endogeneity* (e.g., issues of reverse causality).

The relative sentiment score for a given collection of articles is computed as follows. For the summary statistic of a collection of texts T the total number of occurrences of excitement words, $|Excitement|$, and anxiety words, $|Anxiety|$ are counted, and the numbers scaled by the total text size, $Size[T]$. All words in the dictionaries are given equal weight.[2]

The generic methodology is presented in equation 4.1.

$$RSS[T] = \frac{|Excitement| - |Anxiety|}{size[T]} \tag{4.1}$$

This can be computed on any timescale in order to arrive at a time series of any

[2]It is quite possible that different weights could be assigned to different words, given by the degree of expressed emotion. However, it is also possible that such weights make the methodology less able to generalise to new data. It will become apparent later in the chapter that it is a non-trivial task to select an equally weighted list of words.

given frequency. An increase in this relative sentiment score is due to an increase in excitement and/or a decrease in anxiety, the balance between the two emotions is considered the important variable.

The analysis of this chapter is based on US Reuters news data, US RTRS. Thomson Reuters historical news archive from January 1996 through to October 2014 is analysed to derive an economy wide relative sentiment index. This chapter exclusively analyses this index (and its disaggregated word frequencies) to evaluate some of its main properties, as outlined at the start of the chapter. The following sections on the methodology describes how the complete Reuters database is filtered for the US RTRS subset and how the articles are processed and aggregated.

4.1.1 Article Selection

The Reuters database is filtered for the relevant documents to analyse by the use of regular expressions. The selection of US articles can be described formally as follows.

Definition 1. *Filtering*

Let D be the set of all database 'objects' having key, value pairs (example keys are 'title', 'text', 'date', 'tags' etc.), $\wedge$ be logical conjunction and $\in$ be set-inclusion. Let T denote the collection of 'text' fields (e.g., article body texts) within the set of database objects D. Thus,

$$T = \{t[text] | (t \in D)\}.$$

Given T and a text pattern r let the subset of T consisting of those texts containing (matching) the pattern r be denoted by,

$$T[r \subseteq text] = \{t[text] | (t \in D) \wedge (r \subseteq t[text])\}.$$

A text is here treated as an ordered set of characters so that the subset operator $\subseteq$ may formally be used. The above notation is extended by conditioning the texts on multiple properties more generally.

Definition 2. *Conditioning*

For any collection C of potential values (or a given value) for a given property i (e.g. the 'date' property, the 'tags' property etc.) define the conditional collection of texts $T[C]$ by

$$T[i \in C] = \{t[text] | (t \in D) \wedge (t[i] \in C)\}.$$

The extension to two collections C_1 and C_2 is straight forward:

$$T[i_1 \in C_1, i_2 \in C_2] = \{t[text] | (t \in D) \wedge (t[i_1] \in C_1) \wedge (t[i_2] \in C_2)\}.$$

If C is a single value as opposed to a set, the use of relational symbols different from set inclusion $\in$ is clear from the particular context. These definitions allow a formal definition of how to filter for, e.g., English articles written by Reuters in the US, the main text collection used in this chapter. Finally, the set of non-sports and weather articles considered US focused can be written as follows,

$$\begin{aligned} T^{US} := T[&language = \text{`en'}, attribution = \text{`RTRS'}, \\ &\text{`NEW YORK|WASHINGTON'} \subseteq text, \\ &\{SPO, ODD, WEA\} \cap tags = \emptyset], \end{aligned} \tag{4.2}$$

where | is the symbol of logical disjunction for a regular expression.

4.1.2 Article Tokenisation and Word Frequency Aggregation

In order to count the frequency of the words in the emotion dictionaries a simple tokenisation strategy is carried out. Each article is split into a 'bag-of-words' (i.e., an unordered set of words) using the following procedure:

1. Convert the full article into lowercase letters only (to match the lists of lowercase emotion words)
2. Remove each occurrence of quotation marks
3. Replace each non alphabetic character by a single space character
4. Split the text into words wherever there is a sequence of at least one whitespace character (including newlines, tabs and spaces)
5. Remove any remaining whitespace before or after the resulting words

Technically, steps 2-5 are achieved by replacing each match of the regular expression ['‘] by the empty string, and then replacing each match of the regular expression $[\hat{}a - zA - Z]+$ by the space character ' ', and finally splitting the text at each match of the regular expression $\backslash s+$. From the set of remaining words it is possible to count how many matches there are with the two emotion word dictionaries. The relevant anxiety and excitement word counts are aggregated over the articles produced in a given period. This can be coded in the Scala programming language as follows (this is not the most efficient procedure that is actually used, but it gives the same result and is arguably easier to understand):

```
val anxiety : HashSet[String] = ... // this is the set of anxiety words
val excitement : HashSet[String] = ... // this is the set of excitement words
val articles : List[String] = ... // this is the list of articles for a
    given period
val excitementCount : Double = articles.map { article =>
val tokens = article. toLowerCase.replaceAll("""['`]""",
    "").replaceAll("""[^a-zA-Z]+""", " ").split ("""\s+""")
tokens.filter(excitement.contains(_)).length
} .sum
val anxietyCount : Double = articles.map { article =>
val tokens = article. toLowerCase.replaceAll("""['`]""",
    "").replaceAll("""[^a-zA-Z]+""", " ").split ("""\s+""")
tokens.filter(anxiety.contains(_)).length
} .sum
val sentimentScore : Double = (excitementScore ?
    anxietyScore)/articles.length
```

This procedure is executed for each collection of articles corresponding to a given period.[3] The scores are then sorted chronologically into a time series for further analysis. With a slight modification to the above code it is trivial to count the occurrences of each emotion word individually. The output of such a procedure on US RTRS is used as the basis for much of the analysis in the second part of this chapter, the preliminary evaluation of the RSS methodology.

4.2 Preliminary Evaluation

The simple equation defined at the start of the chapter, Equation 4.1 is applied to Reuters news stories originating in Washington and New York as described earlier. Several aspects of how accurately the latent emotions represented by a given list of words can be measured using this methodology will be considered and evaluated across a few common alternative lists.

The definition in Equation 4.1, although very straight forward, still raises several questions. This chapter will explore several aspects of the RSS measure.

1. What ways are there to adjust for the total text size $Size[T]$?
2. What about negation of words? When, for example, the word 'anxious' is negated as 'not anxious'.

[3]For simplicity this code illustrates the procedure when using the number of articles in the denominator of equation 4.1.

3. What emotion words should be counted in each category?

4. Since different words have very different frequency of use, is it important to adjust for this and if so how?

5. What confidence can be held about a particular relative sentiment score, given the choices of words and the sample size?

The final question is perhaps the most important one and is heavily dependent on the target emotions to be measured.

In order to provide a satisfying answer to question three, two performance measures of the accuracy of a particular list of words are defined. *Clarity* is defined as a measure of the strength of the primary signal captured by the words in the list, and *consistency* is defined as the degree to which all words in the list measure the same signal. These measures provide benchmarks for the performance of different emotion lexicons. Furthermore, in order to refine a particular list, and to judge which words are significant, random matrix theory and principal component analysis are applied together with permutation tests. Individual words can be included or excluded to a particular lexicon by the application of such tests.

This section presents some preliminary means by which the questions raised above can potentially be evaluated. There are several technical contributions within the field of sentiment analysis that are presented in this section [79]. Several novel techniques for measuring the performance of a particular emotion lexicon are developed and applied, which relies on the assumption that the words should all be proxy signals for the underlying target emotional signal. Based on this notion, measures to quantify the degree to which this is the case are developed.

Before evaluating the methodology it is first illustrated on the US RTRS database using the excitement and anxiety lexicons in the following short section. The questions listed above will then be discussed in the order of their appearance.

4.2.1 US RTRS RSS

The analysis when run on all Reuters articles published in the New York or Washington offices excluding 'Sport', 'Weather' and 'Human interest' tags, using the number of articles as scaling variable and aggregating with a monthly frequency (i.e. in the denominator of equation 4.1), can be seen in Figure 4.2.1.

Figure 4.2.2 shows the individual emotion statistics, excitement and anxiety, separately.

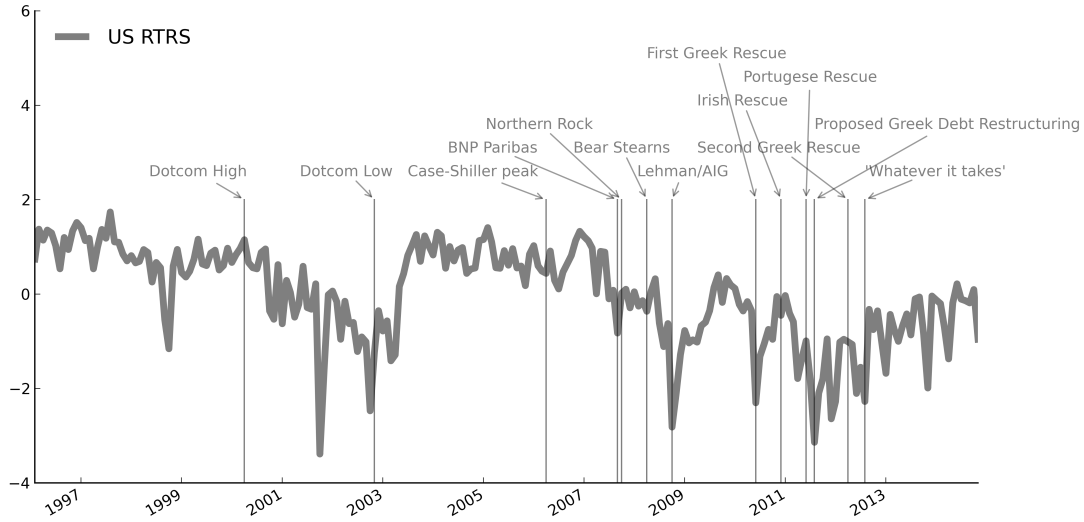

Figure 4.2.1: US RTRS RSS; January 1996 through October 2014

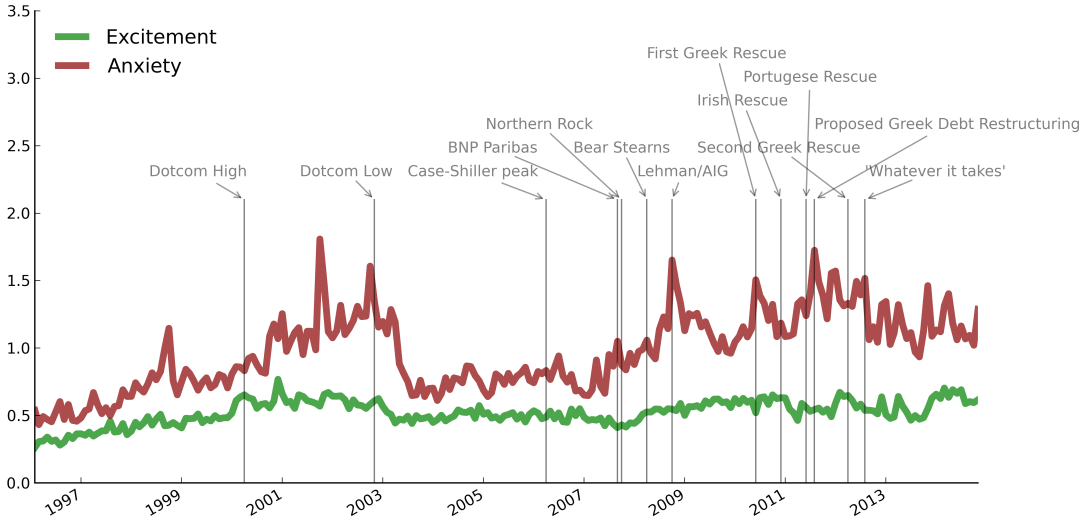

Figure 4.2.2: US RTRS excitement and anxiety components; January 1996 through October 2014

The relative sentiment series follows an intuitive path. Purely for descriptive purposes, the dates of some key economic and financial events are highlighted over the period.[4]

Although the relative difference between the two emotions appear to lack any particular trend, the individual emotions themselves appear to be trending upwards. This raises the question whether or not the emotion words tend to become more frequent over time or if the database is trending in some other way. This, and other properties, will be explored in detail in the following sections.

4.2.2 Scaling $Size[T]$

There are several different ways to measure the total text size. In recent finance and economics literature, e.g. [65], [102], it is common to use the number of words or documents as a scaling variable. However, it is easy to imagine other variables, such as the number of characters or the frequency of common function words such as 'the' or 'and'. The effect of different scaling methods is explored in order to select the one most appropriate for the analysed data sources. It is found that scaling by the number of documents can at times be misleading and that using the number of words or characters might be preferred.

To make sure that shifts in the RSS measure are independent of the total amount of text produced at the time, the effects of different strategies are tested. In other words, a range of common alternative measures of $size[T]$ are explored, such as the number of documents, the number of words and the number of characters, and for comparative purposes also the frequency of the word 'the'.

For example, if it is believed that a document is the most relevant unit of information, i.e., that documents of different lengths should be treated equally even though a priori a longer document might be expected to contain more emotion words, $size[T]$ will be given by the total number of documents. Such a definition might be appropriate if some documents contain more or less 'irrelevant' content than others, for example in the form of numerical tables. However, such a strategy might be slightly misleading if the average length of the documents show trending patterns over time. However, even if such a trend is present for the two emotion lexicons individually, since the variable of interest is the difference between two emotions such trends might be less problematic than first expected.

[4]The timeline of events is partly taken from the annex of the 2009 Financial Stability Report by the Bank of England and partly suggested by Sujit Kapadia (Bank of England) [77]. In other words, the events displayed on the chart were selected independently of the US RTRS RSS index.

Table 4.2.1: Correlations between differently scaled US RTRS excitement series

	WORDS	CHARS	THE	DOCS
WORDS	1	0.996	0.825	0.306
CHARS		1	0.830	0.285
THE			1	0.412
DOCS				1

Table 4.2.2: Correlations between differently scaled US RTRS anxiety series

	WORDS	CHARS	THE	DOCS
WORDS	1	0.999	0.970	0.882
CHARS		1	0.971	0.876
THE			1	0.872
DOCS				1

Table 4.2.3: Correlations between differently scaled US RTRS RSS series

	WORDS	CHARS	THE	DOCS
WORDS	1	1.000	0.954	0.969
CHARS		1	0.955	0.968
THE			1	0.932
DOCS				1

Tables 4.2.1, 4.2.2 and 4.2.3 show the Pearson correlation coefficient between differently scaled series (excitement, anxiety and RSS respectively). WORDS denotes scaling by the total number of words present in the given period. CHARS denotes scaling by the total number of characters present in the given period. THE denotes scaling by the total number of times the word 'the' appears in the given period. Finally, DOCS denotes scaling by the total number of documents present in the given period.

Note that scaling by the number of characters or words gives correlations close to 1.0 for all three series. For RSS, different scaling strategies have a less pronounced effect, yet a noticeable one. The a priori expectation, that possible scaling effects will likely disappear once the difference series (RSS) is constructed, seems consistent with evidence.

When deciding which strategy to apply preference will be given to the one involving the least number of assumptions about the data but still extracts relative sentiment from it as expected. To scale the RSS score by the number of documents would assume that long and short documents should have equal weight. This assumption does not matter if the average length of a document does not change over time in any systematic way. To scale the RSS score by the total number of words would assume that the length of a document is relevant and that the important variable is the 'density' of emotion words in a document (and therefore also in the total collection of documents). This measure is identical to a scaling strategy using the number of documents if the average number of words per document is constant over time (otherwise, the two will yield slightly different results). If the RSS score is scaled by the total number of characters a similar assumption is being made about the length of a document. Since there is no reason to expect that the average length of words would change over time, it is reasonable to believe this method of scaling should yield very similar results to strategy WORDS. If the RSS score is scaled by the total number of times the word 'the' occurs it would assume that the word 'the' has no emotional meaning, but that it could be a good proxy for the total length of actual texts (potentially removing biases due to existence of tables and other types of texts not of a narrative nature). In this case, it is once more assumed that the length of texts matter. In other words, strategy THE has three assumptions, that the word 'the' lacks emotional meaning, that this word count is a good proxy for the length of a text and that the length of text matters. Strategy DOCS has only one assumption, that the length of a document doesn't matter. Strategies WORDS and CHARS have two assumptions, that the length of a document matters and that it is possible to accurately measure it (the second assumption does not seem very significant). Therefore strategy DOCS would be selected in the case of a clear 'tie break' between

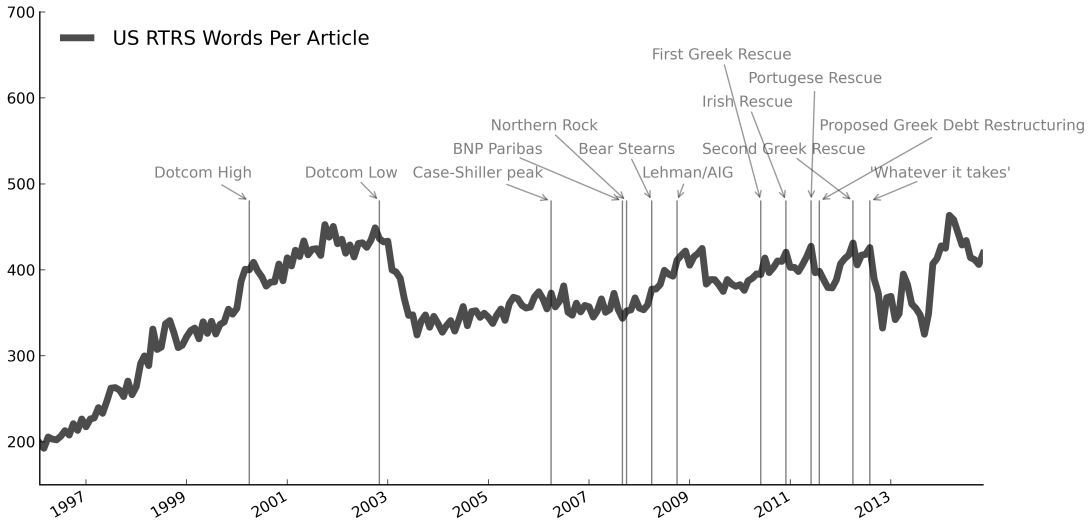

Figure 4.2.3: US RTRS average number of words per article; January 1996 through October 2014

DOCS and WORDS or CHARS or THE. Similarly WORDS or CHARS would be preferred to THE in the presence of a clear tie break.

Since the choice of scaling strategy heavily depends on whether or not the average length of a document depends on time the average length of a document in the US news database is computed. Figure 4.2.3 displays the average number of words per document over time.

It is clear from the chart that the average length of documents increase over the period since 1996. Strategy DOCS is therefore likely to yield significantly different results to the other strategies.

The correlation between the excitement and anxiety series for strategy DOCS is 0.671, for strategy WORDS is -0.145, for strategy CHARS is -0.126 and for strategy THE is 0.134. Closer inspection of the individual emotion series reveals that they both seem to have positive trends, showing that more emotion words are indeed found because the articles tend to be longer. This explains the strong positive correlation in the case of DOCS. It is not clear as to the reason for the positive correlation in the case of THE.

In general it would be expected that the excitement and anxiety series are somewhat negatively correlated. However, it is not expected that they are strictly negatively correlated at all times. There are at least two reasons for this. Recall that the index is constructed by averaging the emotion within all different narratives present in the news database. It is entirely possible that some narratives are highly exciting and others highly anxious during the same period(s). Secondly, it is even

possible that a given narrative can express both high levels of excitement and anxiety in the same period (there could, for example, be a large emotional divide between different social groups). Nonetheless, over the full data series it seems plausible to expect that the two emotions are negatively correlated as opposed to positively correlated if there are aggregate trends in emotion. Thus, if it is generally believed that the two series should be negatively correlated strategies WORDS or CHARS should be preferred over DOCS and THE.

Strategy WORDS is chosen over CHARS as the scaling strategy of choice simply because it will yield a more intuitive interpretation of the final index. However, this choice is inherently *ex post* since it relies on an analysis of the full series. It is quite possible that, e.g., DOCS would have been the selected strategy *a priori*.

4.2.3 Negation

A common criticism of the simple bag-of-words word-count methodology is the presence of negation words that alters the meaning of the target words. Words such as "no", "not", "none", "neither", "never" and "nobody" are commonly used to negate other words [65]. As a robustness check, simple negation detection is implemented which relies on excluding any word from the total word count if it is preceded by either of these words within a window of 3 words. More complex procedures could of course be implemented. However, if this simpler negation detection method (which is expected to cover many cases of negations in practice) is found to have no effect on the final index, it is unlikely that a slightly more sophisticated method would yield a substantially different result.

The index constructed by implementing this simple negation scheme remains correlated with the original US series at 0.999 in both levelled and differenced form. In other words, the effect of negation is simply to change the levels of the emotion series. Negation, as implemented using this procedure, has essentially no effect on the changes in the series. At first this might appear counterintuitive, but in fact a significant difference due to the presence of negation would only be expected if some word (or words) in the dictionary would be more likely negated during only some of the periods considered. In other words, if there would be a systematic bias. Unless such words are present in the emotion dictionaries, for which negation at times becomes more fashionable, the effect over time will cancel out.

Henceforth, the focus of the evaluation will no longer be on the RSS index itself, but rather on the disaggregated word frequencies for a number of different emotion lexicons.

4.2.4 Evaluating Emotion Lexicons

There are several available emotion lexicons. A commonly used source of lexicons can be found in use together with the General Inquirer content analysis system [2]. There are lists of positive and negative words, as well as other types of emotional words. SentiWordNet [7] is a language resource that assigns positivity, negativity and objectivity scores to synonym sets defined by the WordNet resource. Loughran and McDonald [65] defined various lists claiming to be more applicable in a financial context than many of these other lists. This section will compare positive and negative word lists used by the General Inquirer, so called Harvard IV lists, and by Loughran and McDonald, as well as lists of excitement words and anxiety words derived by staff at the *Centre for the Study of Decision-Making Uncertainty* at University College London (for the purpose of operationalising CNT). In the latter case, the words were intuitively selected by an expert psychoanalyst before being validated in a separate study by asking human subjects to rate the words on an excitement and anxiety scale [103]. All lists include different forms and inflections of a word.

Judging whether or not a given emotion lexicon performs well without a quantifiable task (such as classification accuracy given available class labels), or even as expected, is non-trivial. Two main issues are as follows,

- Do the emotion words measure the intended emotions, and nothing else? Put differently, could the words refer to non-emotive concepts (or emotions different from the intended target) such as events, places, times or other objects (which would unintentionally bias the measure or introduce problems of endogeneity)?
- To what extent do the emotion words represent an unbiased sample of all words that could be used to measure the intended emotions?

To evaluate candidate lexicons based on the first consideration, a measure of how much the selected emotion words share the same usage patterns will be applied. If the words where to reflect the assumptions of CNT, i.e., that they are signals for the underlying emotions, it would be expected that the words are used in similar emotional contexts (documents) and therefore have similar temporal frequency paths. On the other hand, if some of the words in the list have usage patterns that are distinctly different from that of the other words, it is much more likely that these words are used to refer to non-emotive concepts (for example to refer to some specific event or object). This simple insight is used to evaluate the internal consistency of a given emotion lexicon, and therefore to what extent the list is 'distinct' to the narrative content.

The same insight can be used to think about the second consideration. If the words in a list appear to share a common pattern it is much likely that the sample is unbiased. A bootstrap procedure can be applied to quantify the presence of a sample bias for any given list.

Through an analysis of the eigenstates of the word frequency time series correlation matrix, ways to quantify how well a particular list of words measure a given latent signal are defined. It is found that the excitement and anxiety lexicons are well suited to the task of testing CNT, with the anxiety lexicon performing particularly well. More specifically, for the purpose of measuring the two emotion groups of approach and avoidance (excitement and anxiety), independently of context (e.g., independently of time, place and events) all words in a particular list would be expected to share common usage patterns, i.e. to correlate with the corresponding emotion. This idea is now made operational by examining to what extent the words in a list show similar patterns over time. Since the emotional change is assumed by CNT to be common across the words in the list, their individual frequencies would be expected to be positively correlated over time. This assumption might be more or less true empirically for different lists, but should be the case for a list suited to measure relative sentiment from the point of view of CNT. However, just because the words in a given list tend to correlate does not preclude the possibility that they are all measures of the wrong concept. Hence, a measure of the degree to which the words of a list are correlated over time can only act as an indication of a necessary, but not sufficient condition, to more suitably measure the emotional groups of interest.

Another, somewhat related, statistical point of interest is the evolution of word frequency ranks over time. How stable are the word ranks between consecutive periods? If there are very large shifts in terms of which words are most frequent, it could imply that some words are more or less dependent on the *content* of what is reported in the news than others. If some words tend to trend more than others, it would likely bias the overall signal away from measuring the intended target, the underlying latent emotional variable.

The excitement and anxiety lists, henceforth referred to as E and A for short, will be compared to the positive and negative word lists created by Loughran and McDonald [65], henceforth referred to as LMP and LMN, positive and negative lexicons from the Harvard-4 lists [2], henceforth referred to as $H4P$ and $H4N$. There are several other lists available, and it is likely that many new lists will be created as the field of sentiment analysis develops, but these are some of the major well established lists that could be considered relevant to evaluate as potentially applicable to CNT.

The following tests will be run on each individual list:

- Using Spearman's rank correlation coefficient, determine the stability of word frequency ranks over time.
- Using random matrix theory and principal component analysis (PCA), identify the words contributing significantly to the principal component. The principal component is likely to be a more reliable measure of the true emotional signal than the individual frequency time series of any particular word. PCA and permutation tests can therefore be applied to determine which words do not positively contribute to this emotional time series signal and this information can be used to quantify the 'coherency' (in the sense that the words measure the evolution of the same emotion) of a particular list. The method could also to be used to refine a particular list of words.

Due to the highly skewed nature of the word frequency distribution, the most frequently occurring words in a list will tend to dominate the overall signal. A long tail of low frequency words would likely be present on the opposite side of the distribution. Therefore, only a collection of the most frequently occurring words in each list will be considered for the above tests.

The 30 most frequent words of each category (positive and negative) are shown in tables 4.2.4 and 4.2.5.

For most of the remaining analysis of this section the top 50 words will be used as they capture most of the information entering the final relative sentiment scores.

Stability Results

The Spearman rank correlation coefficient between word ranks for different periods with a varying number of lags is computed for each word list from January 1996 through October 2014. In particular, the correlations between word ranks in month t and month $t + k$, where k is varied from 1 thorough 24. The Spearman correlation coefficient is used to evaluate how well the relationship of any two variables can be described by a monotonic function. Conceptually, it is the correlation between the relative ranks of variables. By looking at the relative ranks of the words in any given word list, via the Spearman correlations, it is possible to evaluate how stable the relative roles of the words involved are. Results are presented in Tables 4.2.6 and 4.2.7. 95% confidence intervals of the mean correlations are computed for each list by bootstrapping from the samples 10000 times.

Table 4.2.4: The 30 most frequent positive words found in US RTRS; January 1996 through October 2014

E	LMP	H4P
positive	strong	make
boost	gains	rally
great	stable	offset
boosted	good	prime
win	despite	premium
attractive	positive	primarily
enhanced	better	savings
confident	gain	summit
enhancement	highest	pro
encouraging	strength	hand
encourage	gained	pass
encouraged	leading	optional
attract	best	robust
excellent	improvement	upgrade
extraordinary	boost	generate
enhance	stronger	haven
pleased	able	upside
superior	improved	game
attracted	benefit	surge
unique	improve	utilization
impressive	progress	mild
winner	greater	correction
tremendous	favorable	enhancement
promising	effective	supportive
lucrative	great	upbeat
enjoyed	exclusive	allies
enthusiasm	boosted	backing
enjoy	win	boom
winners	improving	bolster
wins	stability	aggregate

Table 4.2.5: The 30 most frequent negative words found in US RTRS; January 1996 through October 2014

A	LMN	H4N
concerns	against	service
negative	late	make
concern	cut	get
fears	loss	run
worries	closed	recession
warned	losses	default
question	decline	charge
uncertainty	concerns	jobless
concerned	negative	try
questions	crisis	volatility
threat	weak	hand
avoid	declined	pass
failure	claims	volatile
warning	lost	exempt
fear	burden	junk
worried	dropped	oversight
rejected	unemployment	serve
stress	problems	antitrust
worry	force	unexpectedly
uncertain	deficit	sluggish
doubt	recession	sour
threatened	sharply	substitution
fail	weaker	exit
warnings	default	risky
failing	closing	slaughter
stressed	concern	mind
questioned	shut	veto
threats	bankruptcy	temporarily
nervous	weakness	ax
doubts	failed	spill

Table 4.2.6: Summary statistic of Spearman rank correlations between word ranks in month t and month $t + k$ for the positive word lists; k is varied from 1 through 24

Statistic	N	Mean	Mean (2.5%,97.5%)
E	5,124	0.813	(0.811, 0.815)
LMP	5,124	0.905	(0.903, 0.906)
H4P	5,124	0.825	(0.823, 0.827)

Table 4.2.7: Summary statistic of Spearman rank correlations between word ranks in month t and month $t + k$ for the negative word lists; k is varied from 1 through 24

Statistic	N	Mean	Mean (2.5%,97.5%)
A	5,124	0.867	(0.866, 0.868)
LMN	5,124	0.770	(0.767, 0.773)
H4N	5,124	0.839	(0.837, 0.841)

The summary statistics indicate that LMP has less drift than the other positive lists and that A has less drift than the other negative lists. The narrow confidence intervals show that the mean values are well-determined. However, the averages do not give very satisfying insights into the nature of changes in the word frequency ranks, for example whether or not a drift tend to persist over time or immediately revert. Heat maps of the Spearman correlation coefficients over time and for different lags, for each list, will be plotted as visual aids. Figure 4.2.4 shows the respective correlation coefficients coloured on a scale from 0.3 (white) to 1 (dark blue) for each list. In each subfigure, the x-axis shows variations over time, from February 1996 through October 2014 and the y-axis shows variations in terms of the number of lags, from 1 through 24 months.

The charts in Figure 4.2.4 need some guidance of interpretation. In general, when the number of lags k increases from 1 through 24 months (displayed on the y-axis of each subplot starting at $k = 1$ at the bottom and ending at $k = 24$ at the top) the Spearman rank correlation drops slightly. This would indicate that the ranks of the words tend to persist, as opposed to immediately return to the previous rank, after the ranks have once changed. There are some periods in which most word lists appear to reorder. At such times, note that rank correlations tend to remain low for several periods at the higher values of k, indicating persistence of the new word ranks.

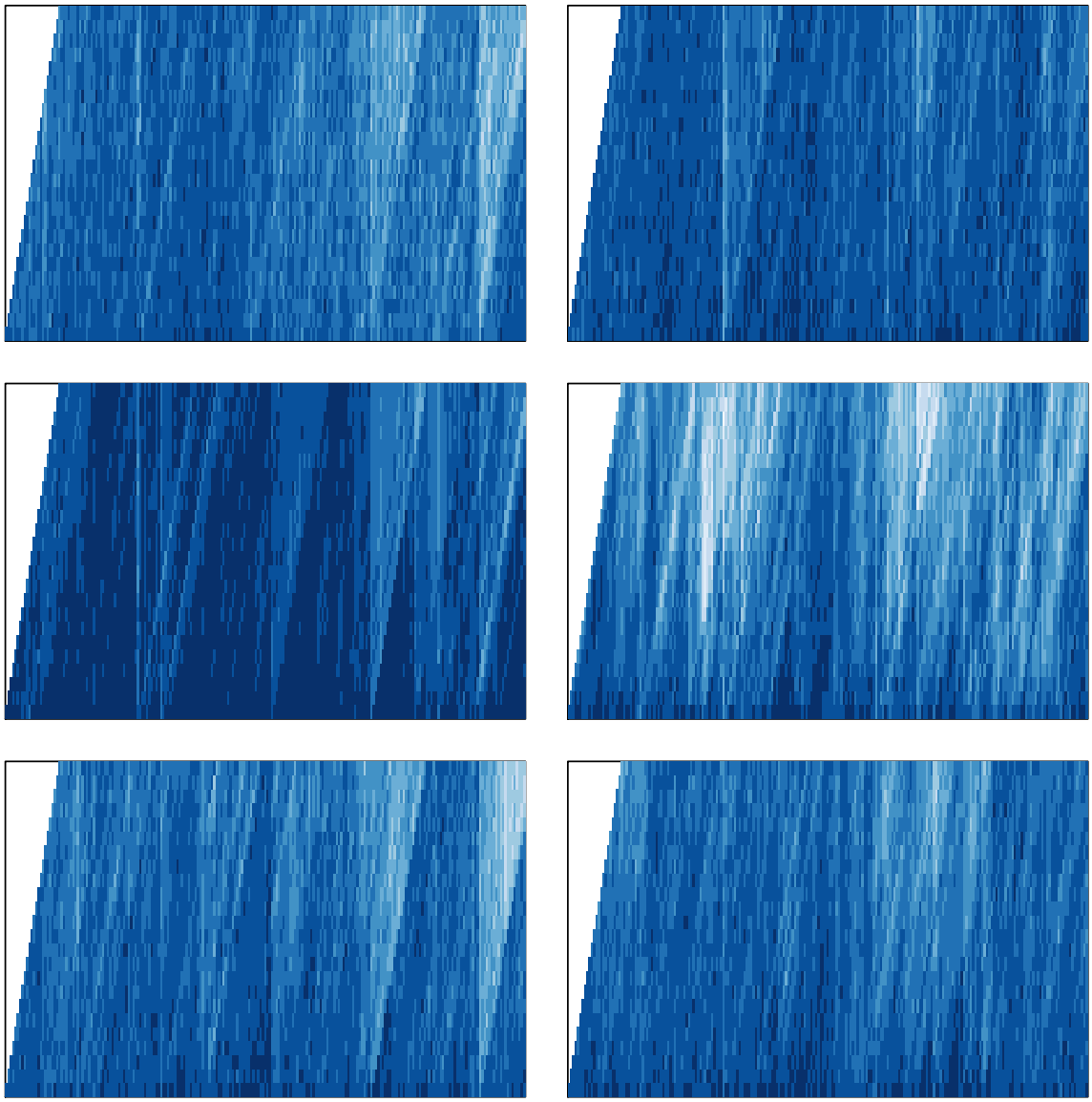

Figure 4.2.4: Spearman rank correlation coefficient over time (x-axis from February 1996 through October 2014) and different lag horizons (y-axis from 1 through 24 months); E (top left), A (top right), LMP (middle left), LMN (middle right), $H4P$ (bottom left) and $H4N$ (bottom right)

Overall, the lists with the highest mean correlation across all periods and lags, A of the 'negative' lists and LMP of the 'positive' lists, appear to experience less drift in the word ranks (indicated by having less pronounced brightly coloured regions). To illustrate the fact that drifts in the relative ranking of the words tend to persist, tables 4.2.8 and 4.2.9 summarise the correlation coefficients over the period for each lag for the lists A and LMP respectively.

Table 4.2.8: Summary statistics of Anxiety rank correlations using the 50 most frequent words; horizon varied from 1 through 24

Horizon	N	Mean	St. Dev.	Min	Max
1	225	0.904	0.035	0.668	0.965
2	224	0.892	0.039	0.681	0.962
3	223	0.891	0.042	0.682	0.951
4	222	0.884	0.041	0.675	0.951
5	221	0.879	0.044	0.637	0.960
6	220	0.878	0.043	0.709	0.959
7	219	0.874	0.042	0.712	0.941
8	218	0.874	0.044	0.706	0.956
9	217	0.877	0.043	0.679	0.972
10	216	0.869	0.043	0.641	0.942
11	215	0.868	0.042	0.721	0.952
12	214	0.872	0.041	0.696	0.962
13	213	0.864	0.046	0.620	0.944
14	212	0.859	0.047	0.638	0.936
15	211	0.862	0.047	0.678	0.955
16	210	0.858	0.048	0.620	0.940
17	209	0.854	0.048	0.668	0.947
18	208	0.852	0.054	0.649	0.945
19	207	0.849	0.051	0.661	0.933
20	206	0.850	0.051	0.658	0.936
21	205	0.851	0.055	0.648	0.941
22	204	0.845	0.057	0.644	0.943
23	203	0.845	0.054	0.682	0.931
24	202	0.848	0.054	0.610	0.953

Table 4.2.9: Summary statistics of LM Positive rank correlations using the 50 most frequent words; horizon varied from 1 through 24

Horizon	N	Mean	St. Dev.	Min	Max
1	225	0.948	0.028	0.821	0.984
2	224	0.938	0.033	0.808	0.981
3	223	0.933	0.039	0.741	0.980
4	222	0.926	0.045	0.667	0.982
5	221	0.923	0.048	0.669	0.982
6	220	0.923	0.045	0.703	0.978
7	219	0.918	0.047	0.671	0.978
8	218	0.917	0.046	0.717	0.976
9	217	0.914	0.048	0.708	0.984
10	216	0.910	0.050	0.699	0.976
11	215	0.908	0.053	0.701	0.975
12	214	0.908	0.053	0.709	0.977
13	213	0.900	0.054	0.715	0.973
14	212	0.896	0.058	0.671	0.972
15	211	0.895	0.057	0.652	0.967
16	210	0.891	0.058	0.679	0.973
17	209	0.888	0.060	0.659	0.974
18	208	0.887	0.061	0.640	0.977
19	207	0.884	0.059	0.677	0.972
20	206	0.879	0.062	0.657	0.969
21	205	0.879	0.064	0.651	0.975
22	204	0.877	0.065	0.657	0.967
23	203	0.876	0.067	0.645	0.970
24	202	0.879	0.063	0.672	0.971

The tables clearly show how the mean rank correlation tends to decline, for both lists, as the horizon is increased from 1 through 24 months.

Clarity and Consistency Results

Time series signals are now derived for each of the top 50 words above by scaling their frequencies by the total number of words, in the same way as done when constructing the main RSS series.

The Pearson correlations between individual normalised word frequency series will indicate the extent to which the frequency tend to move together across the words in the various lists. As described previously, it would be expected that a list which consistently measures a given latent emotional variable has as highly correlated word frequency series as possible. Let M be the $N \times T$ matrix with M_{ij} the relative frequency of word i in period j, $i = 1, \ldots, N$ and $j = 1, \ldots, T$, where $N = 50$ is the number of words and T is the total number of months in the sample. Compute the correlation matrix of M,

$$C = \frac{1}{T} M M^T \tag{4.3}$$

From this matrix, summary statistics of the upper off-diagonal elements are computed for each word list and presented in Table 4.2.10. 95% confidence intervals of the mean correlations are computed for each list by bootstrapping from the samples 10000 times.

Table 4.2.10: Summary statistics of word-by-word correlation matrices for the six word lists

Statistic	N	Mean	Mean (2.5%,97.5%)
E	1,225	0.040	(0.029, 0.051)
A	1,225	0.149	(0.138, 0.161)
LMP	1,225	0.047	(0.029, 0.065)
LMN	1,225	0.081	(0.068, 0.094)
H4P	1,225	−0.003	(-0.017, 0.011)
H4N	1,225	0.007	(-0.006, 0.019)

The last column of table 4.2.10 indicates that the A list is significantly more positively correlated on average than the other lists, followed by $LMNEG$. On the

positive side it is not possible to distinguish between E and $LMPOS$, but both are substantially more positively correlated than both $H4POS$ and $H4NEG$. For both Harvard lists, the average normalised word frequency correlation between the 50 most frequent words cannot be statistically distinguished from zero, as shown by the confidence intervals. In general, the negative lists appear to be more consistent, on average.

However, the summary statistics of the correlation matrices potentially exclude useful information. In particular, whether or not the words tend to cluster. It is possible for the mean correlation of a list to be zero if, for example, the words cluster into two groups which are negatively correlated with each other. To explore the correlation matrices more qualitatively heat map plots of each list are once more created. This is performed for each list, by computing agglomerative clustering of the rows and columns of the corresponding correlation matrix C. The Euclidean distance function is used to cluster the rows and columns. The clustering results are displayed as dendrograms. The rows and columns are sorted within the constraints of the dendrograms to produce the plots, thus revealing potential clusters. The function *heatmap* from the statistical package R is used to produce the plots.

Figures 4.2.6, 4.2.5, 4.2.8, 4.2.7, 4.2.10 and 4.2.9 show heat maps with dendrograms from computing the agglomerative clustering on the Pearson correlation matrices of each word list. The colours range from white, representing a correlation of -1 to dark blue representing a correlation of $+1$.

The correlation charts show an interesting visual pattern. The 'negative' word lists tend to form a single cluster (with the slight exception of $H4N$), whereas the 'positive' word lists tend to form two clusters. By applying concepts from Random Matrix Theory and Principal Component Analysis (PCA) it is possible to determine which words (and therefore how many) account for most of the variability in the principal component. Therefore it is expected that the 'positive' lists have fewer words contributing significantly to the principal 'positive' component than the corresponding number of 'negative' words for the negative component. Empirically, it therefore seems to be a lot of room for improving positive emotion word lists.

The correlation plots can be analysed quantitatively using Random Matrix Theory (RMT) and Principal Component Analysis (PCA), introduced in Chapter 2. The correlation matrix C, defined in equation 4.3, may not be adequately determined when the number of variables (words) relative to the number of observations (in this case, months), over which correlations are computed, is large. This is especially the case when the observations themselves are fairly noisy (this becomes more relevant

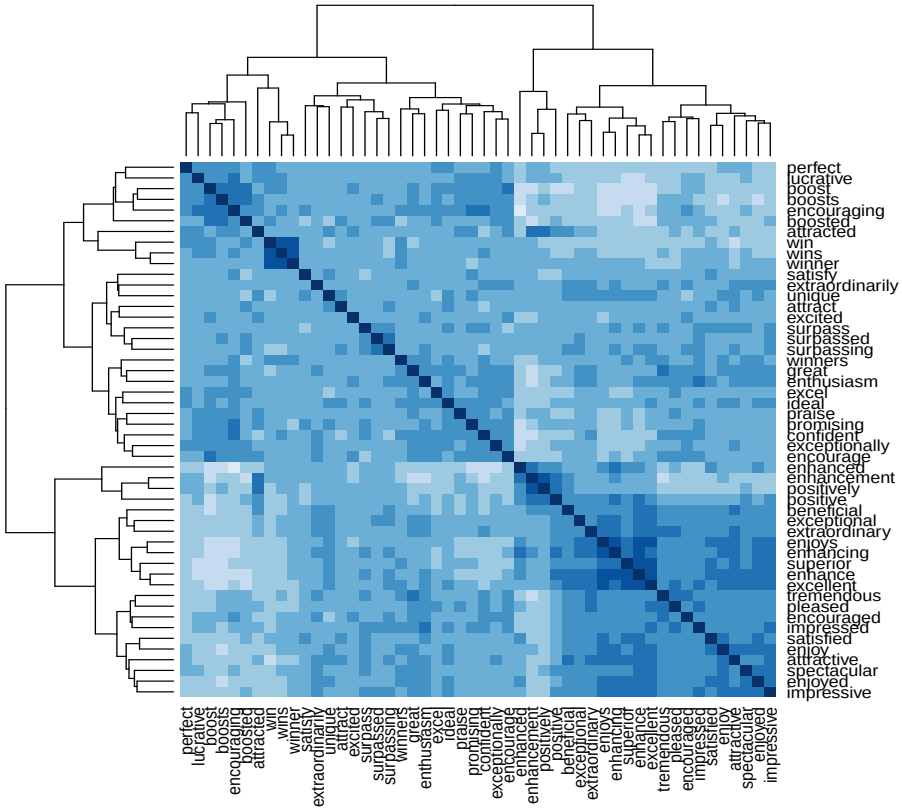

Figure 4.2.5: Correlation matrix of Excitement word frequency series

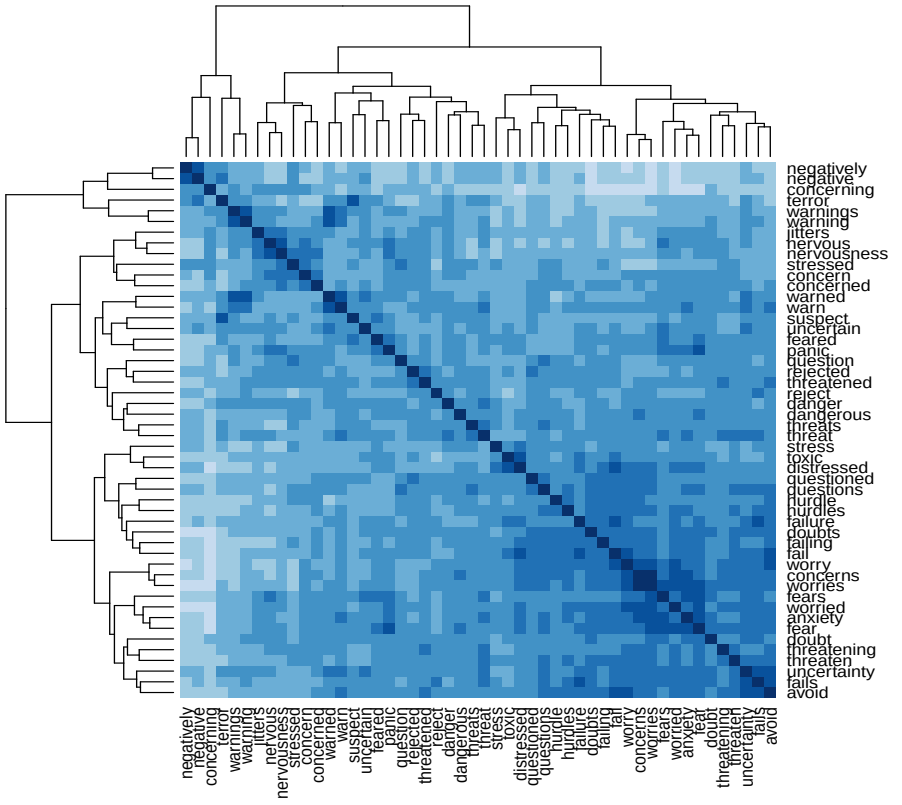

Figure 4.2.6: Correlation matrix of Anxiety word frequency series

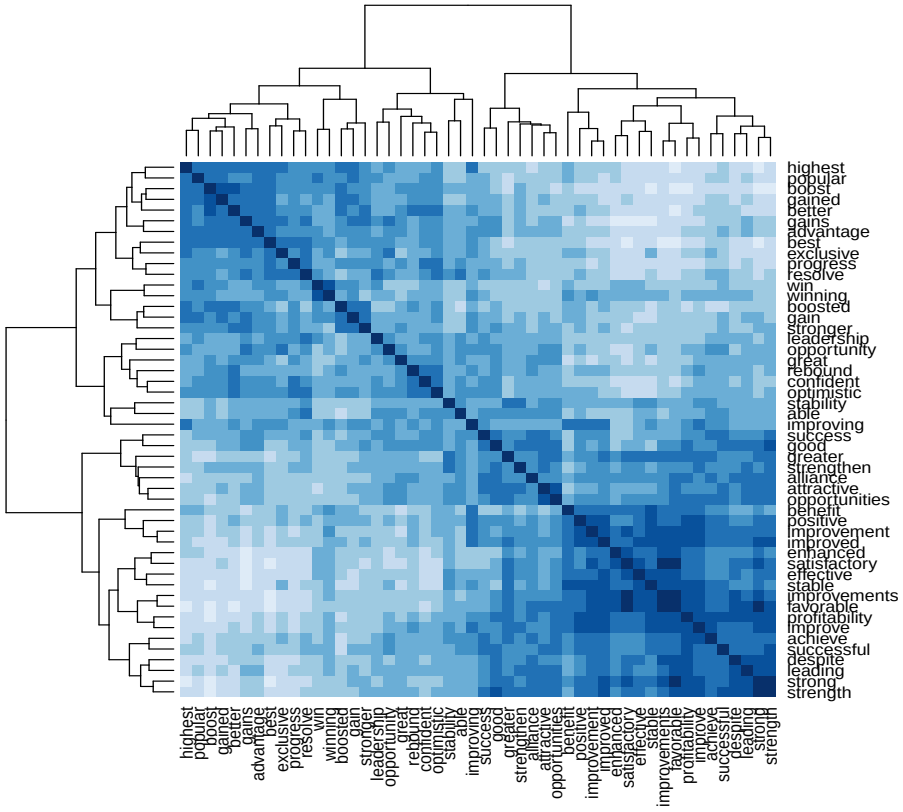

Figure 4.2.7: Correlation matrix of LM Positive word frequency series

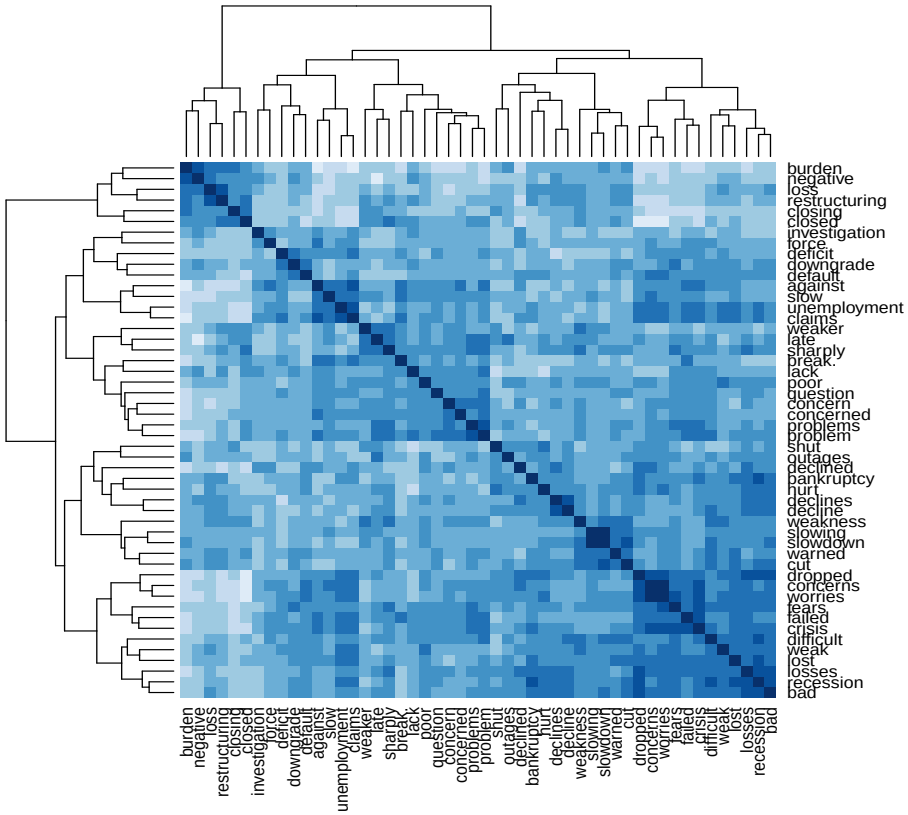

Figure 4.2.8: Correlation matrix of LM Negative word frequency series

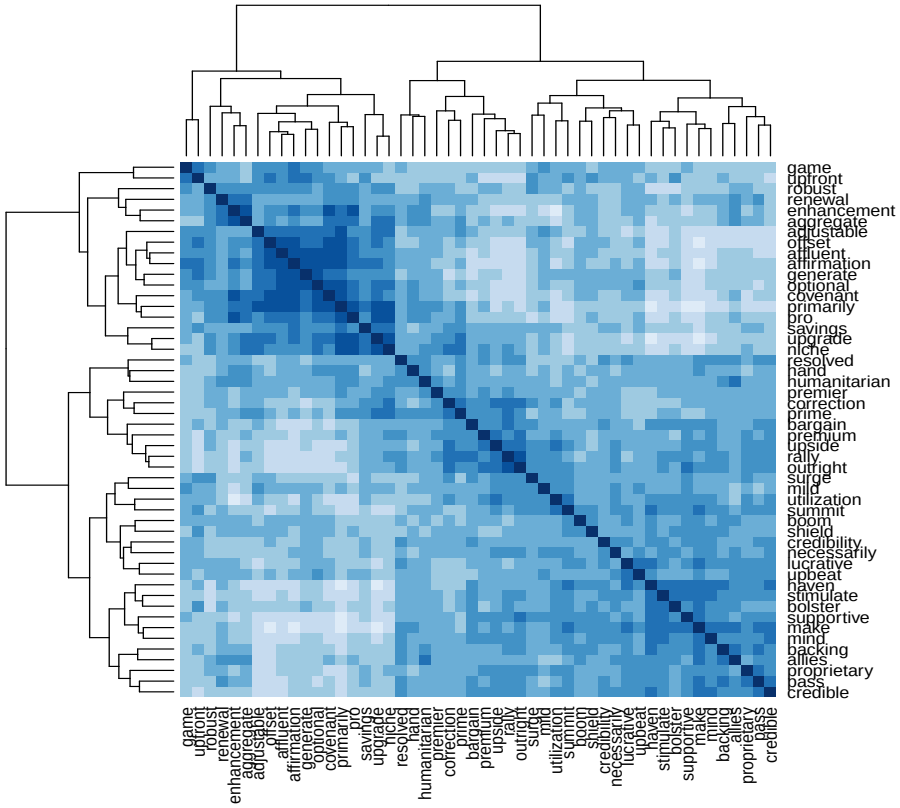

Figure 4.2.9: Correlation matrix of H4 Positive word frequency series

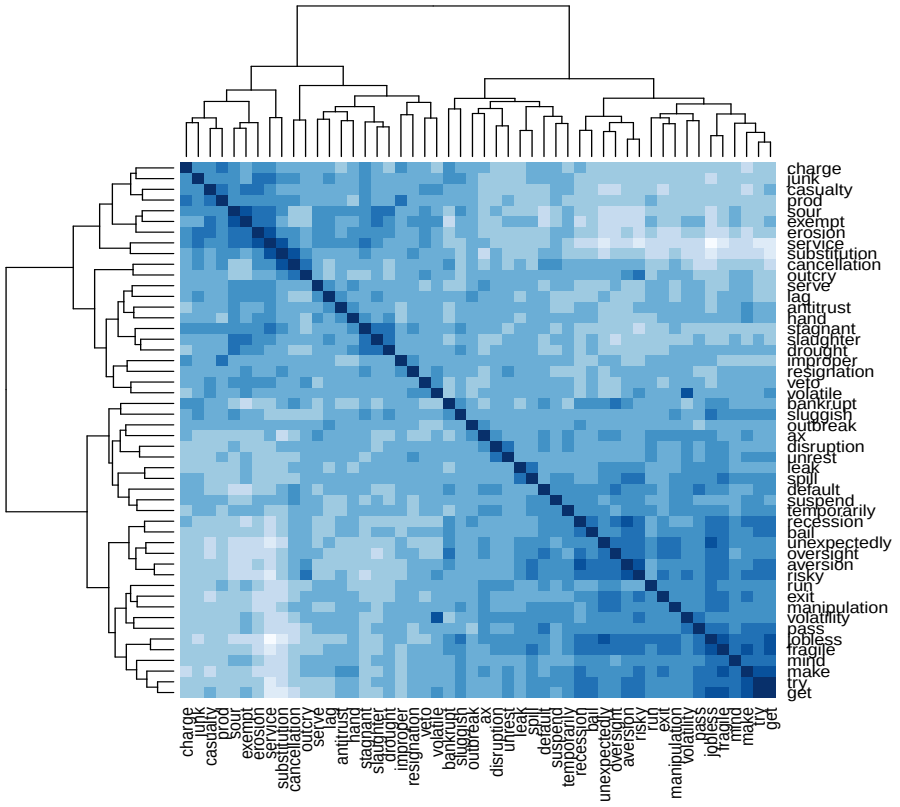

Figure 4.2.10: Correlation matrix of H4 Negative word frequency series

with low frequency words). By application of RMT it is possible to determine the extent of true information in the correlation matrix C, and by using PCA it is possible to determine which words contribute to the true informational signals.

A theorem from Random Matrix Theory concerns the distribution of eigenvalues of random matrices [72]. The fundamental conjecture is re-stated here.

Theorem 2. *Let A be an $N \times N$ real symmetric matrix, whose off-diagonal elements A_{ij} for $i < j$ are independent and identically distributed with zero mean and variance $\sigma > 0$. In the limit $N \to \infty$ the statistical properties of any n eigenvalues of A are independent of the probability density of any element in A.*

For a random correlation matrix, defined above as C, in the limit as $N \to \infty$ and $T \to \infty$ the density of eigenvalues λ is given by [95],

$$p_c(\lambda) = \frac{Q}{2\pi\sigma^2} \frac{\sqrt{(\lambda_{max} - \lambda)(\lambda - \lambda_{min})}}{\lambda} \tag{4.4}$$

for $\lambda \in [\lambda_{min}, \lambda_{max}]$, where $Q = \frac{T}{N} \geq 1$ and,

$$\lambda_{max} = \sigma^2(1 + \frac{1}{\sqrt{Q}})^2 \tag{4.5}$$

$$\lambda_{min} = \sigma^2(1 - \frac{1}{\sqrt{Q}})^2 \tag{4.6}$$

are the upper and lower bounds on the possible eigenvalues. σ^2 is the variance of the elements in the matrix M. Note this means that the individual time series in M must have finite variance for the eigenvalues to have a finite range. A fundamental conjecture on the study of such random matrices postulates that eigenvalues outside this range indicate existence of non-random dependent structure in the correlation matrix. In other words, eigenvalues outside this range would indicate system specific, non-random information. When N and T are finite numbers, the upper and lower bounds, λ_{max} and λ_{min}, will deviate from these theoretical numbers somewhat (in other words, the lower bound can be slightly lower and the upper bound slightly higher than λ_{min} and λ_{max} respectively), and in this case Monte Carlo simulation can be used to determine the exact limits.

The eigenstate of the correlation matrix C with the largest eigenvalue represents the dimension of the data in M with the highest correlations between the words' frequency series. This *principal eigenstate, or principal component*, will be thought of as the primary emotional signal captured by the words in the given list. The words contributing to this component of the correlation matrix can be determined from the corresponding eigenvector. Let e^{max} be the normalised (to be of unit length, measured

in Euclidean distance) eigenvector corresponding to the largest eigenvalue of the correlation matrix C. The entry e_i^{max} represents the (normalised) correlation between word frequency series i and the principal component. The squared entry, $(e_i^{max})^2$, represents the variance of word frequency series i accounted for by the principal component. This *squared component loading* of the PCA solution will be referred to as the *contribution* of word i to the principal component.

In different terms, the eigenvectors of the correlation matrix C correspond to the loadings coefficients of a principal component analysis run on the correlation matrix, as opposed to the co-variance matrix, of the dataset M. The primary interest of this analysis is the principal component of the PCA solution when M is the scaled word frequency time series matrix described above. From the principal component loadings, i.e., the principal eigenvector of the correlation matrix C, the contribution of each word to that component, as defined above, is known. Therefore, define the *target emotion* for a given word list to be this principal component. This definition will assume that the strongest signal captured by the list represents the intended target signal (which may or may not be true but is a reasonable assumption).

The proportion of the variability in the original data that is explained by the principal component (i.e., the target emotion) can be determined by,

$$\text{clarity} = \frac{\lambda_{max}}{\sum_{i=1}^{50} \lambda_i}. \tag{4.7}$$

This provides a measure of the *clarity* with which the target emotion is captured by the words in the list. The larger the principal component, the more uniquely the target emotion can be said to have been captured by the words in the list. This does not specify what the target emotion, i.e. principal component, actually measures. However, the principal component should be the best candidate for the target emotion, under the assumption that all words in a list should measure the same emotional signal.

Furthermore, it is possible to determine the extent to which the words contribute *consistently* to the target emotion by the extent to which the contributions from the different words to the principal component have the same sign. In other words, whether the correlation of a given word frequency series and the principal component is positive or negative. This is easily measured by summing the entries of the principal eigenvector and dividing by the sum of the absolute values of those entries, and disregarding the sign of the result (given that the overall sign of an eigenvector lacks meaning). In other words, define *emotional consistency* of a list by the absolute

value of this ratio,

$$\text{consistency} = \frac{|\sum_{i=1}^{50} e_i^{max}|}{\sum_{i=1}^{50} |e_i^{max}|}, \tag{4.8}$$

where e_i^{max} represents the loading of word i to the principal component (in other words, entry i of the principal eigenvector e^{max} corresponding to the largest eigenvalue λ_{max}).

The two measures, clarity and consistency, can be thought to jointly capture the accuracy of the word lists. On the one hand, the measure of clarity suggests how clearly the target emotion can be distinguished from what is measured by the words. In other words, it suggests to what extent only the target emotion is being measured by the words in the list. On the other hand, the measure of consistency suggests how consistently this target emotion is captured by the different words in the list.

Table 4.2.11 displays the 20 largest eigenvalues of the correlation matrix C for the six word lists.

By applying theorem 2 it is possible to determine the number of significant eigenvalues for each list. For all six lists, $T = 226$ and $N = 50$, and therefore $Q = \frac{T}{N} = \frac{226}{50} = 4.52$. For each list the sample variance has been rescaled to unity, which implies $\lambda_{max} = 2.162$ and $\lambda_{min} = 0.281$. To check whether the sample size substantially changes λ_{max} a Monte Carlo simulation is performed using 10000 random $T \times N$ matrices for which the 50 eigenvalues of each correlation matrix are saved, yielding a total of 500000 empirical eigenvalues. The eigenvalue greater than 99.9% of samples is 2.164, just above the theoretical maximum. Therefore, λ_{max} for each list does indeed carry meaningful information regarding correlated movements of the words.

Table 4.2.12 displays the clarity and consistency scores for the three positive word lists, and Table 4.2.13 displays the scores for the three negative word lists.

In terms of *clarity* of the 'positive' lists, $LMPOS$ comes out on top, followed by $H4POS$ and finally E. However, in terms of the *consistency* score, E greatly outperforms the other two lists (and $H4POS$ is almost entirely inconsistent). This indicates, that although the E list picks up on a weaker target signal, the words in the list tend to contribute equally towards it in a more consistent way than is the case with the other two lists. This suggests the possibility of separating out the principal component of the E list in order to improve on the performance of the overall signal.

In terms of *clarity* of the negative lists, A comes out on top, followed by $H4NEG$ and finally $LMNEG$. In terms of *consistency*, A once more comes out on top as the

Table 4.2.11: Largest eigenvalues of correlation matrices for each of the six wordlists

E	A	LMPOS	LMNEG	H4POS	H4NEG
8.727	11.688	14.587	9.953	11.374	10.509
5.321	4.782	5.626	5.980	4.977	4.055
2.838	3.708	3.719	4.810	3.919	2.681
2.431	2.771	3.341	3.511	2.257	2.565
1.937	2.413	2.244	2.578	2.182	2.264
1.759	2.036	2.064	2.215	1.710	1.880
1.596	1.633	1.546	2.084	1.600	1.793
1.492	1.590	1.274	1.759	1.396	1.578
1.374	1.241	1.125	1.378	1.256	1.408
1.313	1.228	0.975	1.207	1.144	1.354
1.213	1.035	0.894	1.039	1.095	1.288
1.131	0.973	0.751	0.980	1.039	1.184
1.061	0.911	0.723	0.855	1.010	1.172
0.997	0.884	0.674	0.829	0.955	1.039
0.964	0.816	0.666	0.814	0.927	0.972
0.928	0.763	0.629	0.707	0.798	0.927
0.851	0.694	0.597	0.627	0.774	0.864
0.815	0.673	0.588	0.621	0.688	0.839
0.805	0.639	0.542	0.589	0.665	0.783
0.771	0.603	0.516	0.496	0.649	0.723

Table 4.2.12: Clarity and Consistency of each positive word list

Metric	E	LMPOS	H4POS
Clarity	0.175	0.292	0.227
Consistency	0.399	0.254	0.018

Table 4.2.13: Clarity and Consistency of each negative word list

Metric	A	LMNEG	H4NEG
Clarity	0.234	0.199	0.210
Consistency	0.871	0.705	0.271

most consistent of all lists.

Selecting Words Based on Consistency

Turning now to the question of which words contribute significantly to the target emotion given by the principal component. A permutation test will be performed to estimate p-values of the null hypothesis that a given word is independent of the principal component given the structure of the other words in the list, as reported in [64]. Put differently, the null hypothesis that none of the variability in the target word series is accounted for by the principal component. If rejecting the null hypothesis there would be evidence for the alternative hypothesis, that the variability of the given word frequency series can be significantly explained by the principal component, i.e. the target emotion. Performing this hypothesis test on each word in a given list can suggest which words actually contribute significantly to the target emotion. However, with such exploratory research the p-values should be considered with care. Since as many as 50 hypothesis tests are performed for each list one may wish to adjust the significance levels accordingly. In Chapter 2 different p-value adjustment strategies were discussed. Whether or not to adjust depends on the nature of the research, in particular on whether or not one would prefer to perform a Type I error or a Type II error. In other words, to incorrectly conclude that a word significantly contributes to the target emotion or to incorrectly conclude that it does not. Since the aim of this study is to compare the various lists, the choice of whether or not to adjust the significance levels is not all that important since it, a priori, likely effects all lists equally. However, if the result of the analysis was to be used to calibrate a given list to the target database, the question of adjustment becomes very important. Results for each list using uncorrected levels and Bonferroni corrected levels will be reported.

The suggested procedure in [64] is used to test for the contribution of a variable on a specific component of the PCA result. In particular, to test for the significance of the contribution of word i to the principal component given by the eigenvector e^{max} the following steps are performed.

- Decide on the number of permutations to use, N. Repeat the following steps N times
- Given the data matrix M with words as columns and time observations as rows, permute the column of the target word i while keeping the other columns fixed. Call the new matrix $MPerm$
- Compute a PCA using the correlation matrix of $MPerm$

- Save the squared loading of the principal component of the PCA result for the target word (i.e., the corresponding squared entry of the principal eigenvector, $(e_i^{max})^2$)

Given this strategy it is possible to say that the squared loading of the target word i on the principal component of the observed PCA model computed on M is significant if the p-value, as defined by,

$$p = \frac{q+1}{N+1} \tag{4.9}$$

where q is the number of times the squared loading from the permutation distribution generated according to the above procedure exceeds the squared loading of the observed PCA model, is less than or equal to a specified significance level. A biased estimator of the p-value is used as opposed to the unbiased estimator $p = \frac{q}{N}$ because the entire permutation space is not enumerated, only samples (with replacement) from the space of possible permutations. An unbiased estimator of the p-value would be inappropriate and yield more Type I errors than assumed due to the uncertainty inherent in the estimation process. It is shown in [100] that the biased estimator $p = \frac{q+1}{N+1}$ overestimates the exact p-value owing to the effect of sampling, due to the small probability that the observed data are sampled and therefore included in the permutation distribution. If sampling without replacement the biased estimator above would be exact. When the total number of possible permutations is equal to 1000, a p-value of 0.05 would be approximately overestimated by 2%. In general, the p-value used is overestimated by

$$\int_0^{\frac{0.5}{N_t+1}} F(q; N, p_t) dp_t,$$

where N_t is the total number of possible permutations of a word frequency series, $F(q; N, p_t)$ is the cumulative probability function of the binomial distribution and $p_t = \frac{q_t+1}{N_t+1}$ is the exhaustive permutation p-value obtained if considering all distinct permutations [100]. Since each word series has 226 time observations (and no repeated entries in the matrix M for any of the lists, since only the most frequent words are considered[5]), N_t will be very large (in fact, N_t is given by the factorial number of 226, which is a very large number) and this problem will not be of much concern. A test will be judged significant if the generated p-value is less than or equal to a significance threshold.

[5]This is confirmed empirically, but it is also easy to imagine that this should be true since the total number of words for each period is a unique number and it is also unlikely for two words in different lists to have the same frequency for a given period.

Note that words with inconsistent contribution to the principal component (in the sense that their contribution have the opposite sign of the sum of all word contributions) are still considered since the test statistic used is the squared component contribution, i.e. the word variance accounted for by the principal component. The significance level α is set to 0.05. Therefore, a simple Bonferroni correction for a total of $50 * 6 = 300$ tests yields a corrected level of $\frac{\alpha}{300} = 0.00017$

In order to have a minimum estimated p-value smaller or equal to the Bonferroni corrected level, N must be chosen to satisfy $\frac{1}{1+N} \leq 0.00017$. In other words, N must be at least 5882. Therefore, let $N = 9999$ which yields a minimum p-value of 0.0001.

The corresponding p-value estimates when running the above procedure with 9999 permutation samples with replacement for each word in each list are reported in Tables 4.2.14 and 4.2.15, for the positive and negative lists respectively.[6]

Table 4.2.18 shows the number of words significantly contributing to the principal component (regardless of the sign of the contribution) with the two different levels of significance. The table shows that a reasonably large number of words tend to contribute significantly to the principal component. However, including the words with a component contribution of the opposite sign as the overall component contribution is misleading since only 'positive' contributions to the principal component are of interest.

Tables 4.2.16 and 4.2.17 display the Pearson r correlation between the principal eigenvectors and the words of the positive and negative lists respectively. Here, the eigenvectors have been multiplied by the sign of the sum of their entries. Hence, a positive entry can be interpreted as corresponding to a word that contributes 'positively' to the principal component.

Table 4.2.19 shows the number of words which significantly contribute to the principal component and for which the sign of that contribution has the same sign as the overall sum of principal component contributions. In other words, all words j for which $sign(e_j^{max}) \neq sign(\sum_i e_i^{max})$ have been excluded.

In this case, only significance tests on the words with the same signed contribution have been performed, as opposed to the full list of 300 words. The Bonferroni corrected significance level is therefore obtained by dividing $\alpha = 0.05$ by the total number of such words, which is 206. In other words, the corrected significance level is

[6]Interestingly, in the E list the word 'positive' has a significant contribution whereas the word 'positively' does not. This illustrates the importance of context and shows that it is not always possible to assume that inflections of the same lemma will necessarily capture the same emotional time series signal, as is a common assumption, e.g., [118].

Table 4.2.14: Permutation p-values for squared contribution to principal component for all positive lists

E	p-value	LMPOS	p-value	H4POS	p-value
positive	1e-04	strong	1e-04	make	1e-04
boost	1e-04	gains	1e-04	rally	1e-04
boosted	1e-04	stable	1e-04	offset	1e-04
win	1e-04	despite	1e-04	primarily	1e-04
attractive	1e-04	positive	1e-04	savings	1e-04
encouraging	1e-04	better	1e-04	summit	1e-04
encouraged	1e-04	gain	1e-04	pro	1e-04
excellent	1e-04	highest	1e-04	pass	1e-04
extraordinary	1e-04	strength	1e-04	optional	1e-04
enhance	1e-04	gained	1e-04	robust	1e-04
pleased	1e-04	leading	1e-04	upgrade	1e-04
superior	1e-04	best	1e-04	generate	1e-04
impressive	1e-04	improvement	1e-04	haven	1e-04
tremendous	1e-04	boost	1e-04	upside	1e-04
lucrative	1e-04	improved	1e-04	utilization	1e-04
enjoyed	1e-04	benefit	1e-04	enhancement	1e-04
enjoy	1e-04	improve	1e-04	supportive	1e-04
perfect	1e-04	progress	1e-04	allies	1e-04
satisfied	1e-04	greater	1e-04	backing	1e-04
boosts	1e-04	favorable	1e-04	bolster	1e-04
beneficial	1e-04	effective	1e-04	aggregate	1e-04
exceptional	1e-04	exclusive	1e-04	mind	1e-04
enjoys	1e-04	boosted	1e-04	affirmation	1e-04
enhancing	1e-04	win	1e-04	adjustable	1e-04
spectacular	1e-04	improvements	1e-04	outright	1e-04
impressed	1e-04	successful	1e-04	bargain	1e-04
enhanced	2e-04	opportunities	1e-04	covenant	1e-04
attracted	2e-04	satisfactory	1e-04	stimulate	1e-04
wins	2e-04	profitability	1e-04	lucrative	1e-04
unique	0.0013	advantage	1e-04	proprietary	1e-04
surpassing	0.0067	attractive	1e-04	niche	1e-04
praise	0.0189	popular	1e-04	upfront	1e-04
exceptionally	0.0192	alliance	1e-04	credible	1e-04
promising	0.0272	enhanced	1e-04	affluent	1e-04
surpass	0.0288	achieve	1e-04	premium	2e-04
extraordinarily	0.0316	optimistic	1e-04	game	6e-04
confident	0.0323	confident	1e-04	credibility	6e-04
winner	0.0442	strengthen	1e-04	surge	7e-04
enthusiasm	0.0477	resolve	1e-04	necessarily	7e-04
excel	0.0535	good	2e-04	boom	0.0115
encourage	0.1327	stronger	2e-04	renewal	0.0115
great	0.2821	leadership	6e-04	shield	0.0244
winners	0.2927	great	0.0032	upbeat	0.0727
enhancement	0.3844	opportunity	0.0073	correction	0.0741
excited	0.5009	success	0.009	resolved	0.0769
attract	0.5581	rebound	0.0266	mild	0.1514
ideal	0.7297	stability	0.0296	hand	0.3192
satisfy	0.7627	winning	0.0901	humanitarian	0.6229
surpassed	0.8049	improving	0.5139	prime	0.6871
positively	0.8063	able	0.9718	premier	0.8627

Table 4.2.15: Permutation p-values for squared contribution to principal component for all negative lists

A	p-value	LMNEG	p-value	H4NEG	p-value
concerns	1e-04	against	1e-04	service	1e-04
negative	1e-04	closed	1e-04	make	1e-04
fears	1e-04	losses	1e-04	get	1e-04
worries	1e-04	concerns	1e-04	run	1e-04
question	1e-04	negative	1e-04	recession	1e-04
uncertainty	1e-04	crisis	1e-04	default	1e-04
questions	1e-04	weak	1e-04	charge	1e-04
threat	1e-04	declined	1e-04	jobless	1e-04
avoid	1e-04	claims	1e-04	try	1e-04
failure	1e-04	lost	1e-04	volatility	1e-04
fear	1e-04	burden	1e-04	pass	1e-04
worried	1e-04	dropped	1e-04	exempt	1e-04
rejected	1e-04	unemployment	1e-04	junk	1e-04
worry	1e-04	problems	1e-04	oversight	1e-04
uncertain	1e-04	recession	1e-04	unexpectedly	1e-04
doubt	1e-04	default	1e-04	sour	1e-04
threatened	1e-04	closing	1e-04	substitution	1e-04
fail	1e-04	bankruptcy	1e-04	exit	1e-04
failing	1e-04	failed	1e-04	risky	1e-04
questioned	1e-04	fears	1e-04	slaughter	1e-04
threats	1e-04	slow	1e-04	mind	1e-04
doubts	1e-04	worries	1e-04	temporarily	1e-04
dangerous	1e-04	problem	1e-04	spill	1e-04
feared	1e-04	hurt	1e-04	casualty	1e-04
fails	1e-04	difficult	1e-04	fragile	1e-04
distressed	1e-04	restructuring	1e-04	bail	1e-04
danger	1e-04	bad	1e-04	aversion	1e-04
threatening	1e-04	poor	1e-04	erosion	1e-04
threaten	1e-04	sharply	2e-04	prod	1e-04
negatively	1e-04	question	2e-04	manipulation	1e-04
anxiety	1e-04	concerned	2e-04	stagnant	1e-04
concerning	1e-04	concern	3e-04	serve	2e-04
panic	1e-04	declines	3e-04	leak	5e-04
warn	1e-04	cut	4e-04	cancellation	0.0015
toxic	1e-04	loss	4e-04	ax	0.0016
hurdles	1e-04	decline	4e-04	lag	0.0022
hurdle	1e-04	slowing	4e-04	suspend	0.0036
concerned	3e-04	slowdown	6e-04	unrest	0.0059
stress	4e-04	deficit	0.0011	drought	0.0061
warned	5e-04	warned	0.0012	bankrupt	0.0089
concern	7e-04	late	0.0053	disruption	0.0118
reject	0.0021	weakness	0.0117	veto	0.0656
suspect	0.015	outages	0.0143	resignation	0.0686
nervousness	0.1399	force	0.038	hand	0.0718
nervous	0.2029	downgrade	0.0555	sluggish	0.0784
jitters	0.2419	break.	0.1305	improper	0.1225
terror	0.2509	investigation	0.2426	antitrust	0.1705
stressed	0.322	weaker	0.3863	outcry	0.6119
warning	0.3786	lack	0.508	outbreak	0.9242
warnings	0.5209	shut	0.7963	volatile	0.9636

Table 4.2.16: Pearson correlation between principal component and words of each positive list

E	Norm. Pearson r	LMPOS	Norm. Pearson r	H4POS	Norm. Pearson r
enhance	0.29	favorable	0.24	make	0.21
excellent	0.273	strong	0.223	haven	0.199
enjoys	0.257	profitability	0.222	supportive	0.195
enhancing	0.244	improvements	0.22	mind	0.176
superior	0.226	improved	0.2	stimulate	0.169
impressive	0.206	improve	0.197	outright	0.152
enjoyed	0.204	satisfactory	0.197	credible	0.149
attractive	0.195	strength	0.195	backing	0.146
extraordinary	0.191	effective	0.193	pass	0.145
enjoy	0.19	stable	0.188	summit	0.139
beneficial	0.176	improvement	0.185	upside	0.134
exceptional	0.166	despite	0.173	rally	0.132
spectacular	0.165	leading	0.161	utilization	0.129
satisfied	0.162	achieve	0.158	bolster	0.128
tremendous	0.141	enhanced	0.157	proprietary	0.121
impressed	0.136	successful	0.139	bargain	0.111
pleased	0.135	positive	0.139	allies	0.086
encouraged	0.113	greater	0.132	premium	0.085
positive	0.112	opportunities	0.124	lucrative	0.083
enhanced	0.098	alliance	0.117	credibility	0.076
unique	0.081	benefit	0.103	surge	0.075
surpassing	0.068	attractive	0.091	necessarily	0.072
surpass	0.055	strengthen	0.087	boom	0.05
extraordinarily	0.055	good	0.077	shield	0.049
enthusiasm	0.05	success	0.049	correction	0.039
great	0.028	stability	0.04	upbeat	0.039
winners	0.027	improving	0.012	resolved	0.038
enhancement	0.022	able	0.001	mild	0.031
attract	0.015	winning	-0.032	premier	0.004
ideal	0.009	rebound	-0.041	prime	-0.009
satisfy	0.008	opportunity	-0.051	humanitarian	-0.011
surpassed	0.006	great	-0.054	hand	-0.022
positively	-0.006	leadership	-0.065	renewal	-0.053
excited	-0.017	stronger	-0.07	game	-0.069
encourage	-0.039	optimistic	-0.081	upfront	-0.08
excel	-0.049	win	-0.083	savings	-0.11
winner	-0.051	confident	-0.09	robust	-0.12
confident	-0.054	gain	-0.094	aggregate	-0.12
promising	-0.057	boosted	-0.103	niche	-0.158
exceptionally	-0.059	resolve	-0.106	enhancement	-0.168
praise	-0.06	exclusive	-0.119	generate	-0.177
wins	-0.095	advantage	-0.123	adjustable	-0.187
attracted	-0.096	progress	-0.133	upgrade	-0.194
perfect	-0.127	popular	-0.145	optional	-0.197
boosted	-0.138	better	-0.151	pro	-0.205
win	-0.146	gained	-0.152	affluent	-0.215
lucrative	-0.16	gains	-0.154	offset	-0.227
encouraging	-0.171	highest	-0.172	covenant	-0.23
boosts	-0.213	best	-0.181	affirmation	-0.232
boost	-0.223	boost	-0.19	primarily	-0.265

Table 4.2.17: Pearson correlation between principal component and words of each negative list

A	Norm. Pearson r	LMNEG	Norm. Pearson r	H4NEG	Norm. Pearson r
fear	0.231	crisis	0.253	jobless	0.245
worries	0.225	worries	0.247	fragile	0.231
worry	0.224	dropped	0.247	risky	0.219
concerns	0.217	concerns	0.237	get	0.213
anxiety	0.215	losses	0.23	unexpectedly	0.211
fail	0.21	bad	0.218	try	0.204
worried	0.209	recession	0.217	make	0.203
avoid	0.204	fears	0.212	aversion	0.197
uncertainty	0.192	lost	0.2	oversight	0.194
doubts	0.191	failed	0.2	exit	0.183
fears	0.186	claims	0.189	bail	0.173
threaten	0.18	difficult	0.185	recession	0.171
fails	0.18	unemployment	0.179	pass	0.161
questions	0.176	weak	0.172	manipulation	0.153
failing	0.172	problem	0.142	mind	0.139
hurdles	0.16	against	0.136	volatility	0.135
doubt	0.159	slow	0.118	default	0.118
threatening	0.157	bankruptcy	0.117	spill	0.115
failure	0.154	problems	0.115	run	0.114
distressed	0.149	hurt	0.107	temporarily	0.096
questioned	0.142	poor	0.103	leak	0.075
threat	0.132	default	0.102	ax	0.069
hurdle	0.128	question	0.097	suspend	0.065
warn	0.124	declined	0.096	unrest	0.061
panic	0.123	cut	0.092	bankrupt	0.06
dangerous	0.121	sharply	0.09	disruption	0.056
feared	0.119	concerned	0.088	sluggish	0.04
threats	0.117	slowing	0.083	outcry	0.012
threatened	0.112	declines	0.082	outbreak	0.002
toxic	0.106	concern	0.082	volatile	0.001
danger	0.099	slowdown	0.081	antitrust	-0.031
question	0.097	decline	0.08	improper	-0.035
concerned	0.088	warned	0.078	hand	-0.041
rejected	0.088	deficit	0.074	resignation	-0.041
uncertain	0.086	late	0.064	veto	-0.042
stress	0.086	weakness	0.059	drought	-0.061
warned	0.077	outages	0.057	lag	-0.068
concern	0.069	force	0.049	cancellation	-0.072
reject	0.065	downgrade	0.044	serve	-0.083
suspect	0.051	break.	0.036	slaughter	-0.111
nervousness	0.031	lack	0.016	charge	-0.115
nervous	0.027	shut	0.006	prod	-0.126
jitters	0.025	weaker	-0.02	junk	-0.133
stressed	0.021	investigation	-0.028	stagnant	-0.134
warning	0.019	loss	-0.081	casualty	-0.153
warnings	-0.014	restructuring	-0.105	exempt	-0.157
terror	-0.025	negative	-0.141	sour	-0.187
negative	-0.112	closed	-0.152	substitution	-0.188
negatively	-0.119	closing	-0.177	erosion	-0.196
concerning	-0.139	burden	-0.209	service	-0.27

in this case given by $\frac{\alpha}{206} = 0.00024$.

Table 4.2.18: Number of words significantly contributing to the principal components of each word list; significance level $\alpha = 0.05$

Correction	E	A	LMPOS	LMNEG	H4POS	H4NEG
Uncorrected	39	43	47	44	42	41
Bonferroni	26	37	39	28	34	31

Table 4.2.19: Number of words significantly contributing to the principal components of each word list and have an equally signed contribution (correlation with the principal component); significance level $\alpha = 0.05$

Correction	E	A	LMPOS	LMNEG	H4POS	H4NEG
Uncorrected	25	40	26	38	24	26
Bonferroni	20	34	24	26	19	20

The table confirms the hypothesis that the negative lists tend to have more words contributing to the principal component than the positive lists. This turns out to be the case for all three lists and independently of simple significance level correction, in the form of Bonferroni correction. The A list is the most clearly defined of all lists, followed by $LMNEG$, $LMPOS$, $H4NEG$, E and finally $H4POS$.

Finally, given the hypothesis that the principal component of each list more likely corresponds to the intended emotional target variable than either of the individual word frequency series themselves (and as such also the simple aggregate), it would be expected that the correlations between the principal components of the various lists are higher than the correlations between the full aggregate RSS series.

It is found, rather strikingly, that the principal components of each list are much more highly correlated with one another than the original series, as can be seen in Tables 4.2.20 and 4.2.21. This is especially true for the negative lists, where in each case the correlation between the original aggregate series produced with the 50 most frequent words is lower than the corresponding correlation between the respective principal components. This supports the view that the target emotion is likely to be better approximated by the principal component than with the aggregate series.

Table 4.2.20: Correlations between aggregate positive series generated using the 50 most frequent words in each list and the corresponding principal components

	E	LMPOS	H4POS	E.PC1	LMPOS.PC1	H4POS.PC1
E	1	0.646	0.215	0.570	0.183	0.202
LMPOS	0.646	1	0.333	0.527	0.485	-0.219
H4POS	0.215	0.333	1	0.141	0.176	0.058
E.PC1	0.570	0.527	0.141	1	0.802	-0.430
LMPOS.PC1	0.183	0.485	0.176	0.802	1	-0.814
H4POS.PC1	0.202	-0.219	0.058	-0.430	-0.814	1

Table 4.2.21: Correlations between aggregate negative series generated using the 50 most frequent words in each list and the corresponding principal components

	A	LMNEG	H4NEG	A.PC1	LMNEG.PC1	H4NEG.PC1
A	1	0.707	0.760	0.929	0.802	0.690
LMNEG	0.707	1	0.644	0.710	0.876	0.545
H4NEG	0.760	0.644	1	0.764	0.724	0.666
A.PC1	0.929	0.710	0.764	1	0.903	0.869
LMNEG.PC1	0.802	0.876	0.724	0.903	1	0.840
H4NEG.PC1	0.690	0.545	0.666	0.869	0.840	1

An interesting case is $H4POS$ whose principal component is in fact negatively correlated with the principal components of the other positive lists. Recall that the consistency score of this list was as low as 0.018, indicating that this principal component is very badly determined by the list. In fact, the second largest component of $H4POS$ is more consistent with a score of 0.142. In this case, the correlation between $H4POS.PC2$ and E is 0.631, with $LMPOS$ is 0.480, with $H4POS$ is 0.277, with $E.PC1$ is 0.765 and with $LMPOS.PC1$ is 0.435. In other words, the second component of $H4POS$ is likely to be a better representation of what the E and $LMPOS$ lists attempt to measure. In conclusion, the $H4POS$ list is very badly suited to the application of CNT.

By considering the correlation between the original aggregate series and the corresponding principal components it is possible to judge the extent to which the original series approximate the principal component *ex ante*. In other words, without knowledge of the correlational structure over the full period (which was used to construct the PCA model), how well would the aggregate series have captured the emotional target represented by the principal component? The correlation between E and $E.PC1$ at 0.570 is much higher than the correlations between $LMPOS$ and $LMPOS.PC1$ at 0.485 and between $H4POS$ and $H4POS.PC1$ at 0.058 (which again is evidence of how badly determined the $H4POS$ list is). Likewise, in case of the negative lists, the correlation between A and $A.PC1$ at 0.929 is much higher than that between $LMNEG$ and $LMNEG.PC1$ at 0.876 and between $H4NEG$ and $H4NEG.PC1$ at 0.666. Overall, the A list once more proves to be better determined than the other lists.

Tables 4.2.22 and 4.2.23 list the words in each positive and negative list respectively that appear as having a significant contribution to the principal component with an uncorrected significance level, $\alpha = 0.05$.

Given the analysis conducted in this section it seems a reasonable conclusion to say that the excitement and anxiety lexicons are comparatively well defined for the purpose of measuring relative sentiment shifts from the perspective of CNT. The anxiety list can be considered to more accurately capture the intended emotional signal in this manner than the other 'negative' lists (indeed, under the measures of accuracy considered here, it is the most accurate of all lists) selected for comparison. The excitement list could potentially be improved by carefully excluding some of the words, perhaps by first considering the words that did not make it into Table 4.2.22, and possibly including some of the words from the $LMPOS$ list which appears to be one of the more consistent of the positive lists compared in this section.

Table 4.2.22: Positive emotion words that significantly contribute to the principal emotional signal

E Words	E p-values	LMPOS Words	LMPOS p-values	H4POS Words	H4POS p-values
positive	1e-04	strong	1e-04	make	1e-04
attractive	1e-04	stable	1e-04	rally	1e-04
enhanced	2e-04	good	2e-04	premium	2e-04
encouraged	1e-04	despite	1e-04	summit	1e-04
excellent	1e-04	positive	1e-04	pass	1e-04
extraordinary	1e-04	strength	1e-04	haven	1e-04
enhance	1e-04	leading	1e-04	upside	1e-04
pleased	1e-04	improvement	1e-04	surge	7e-04
superior	1e-04	improved	1e-04	utilization	1e-04
unique	0.0013	benefit	1e-04	supportive	1e-04
impressive	1e-04	improve	1e-04	allies	1e-04
tremendous	1e-04	greater	1e-04	backing	1e-04
enjoyed	1e-04	favorable	1e-04	boom	0.0115
enthusiasm	0.0477	effective	1e-04	bolster	1e-04
enjoy	1e-04	stability	0.0296	necessarily	7e-04
satisfied	1e-04	improvements	1e-04	mind	1e-04
beneficial	1e-04	successful	1e-04	outright	1e-04
exceptional	1e-04	opportunities	1e-04	bargain	1e-04
enjoys	1e-04	satisfactory	1e-04	credibility	6e-04
enhancing	1e-04	profitability	1e-04	stimulate	1e-04
surpassing	0.0067	attractive	1e-04	lucrative	1e-04
spectacular	1e-04	success	0.009	proprietary	1e-04
surpass	0.0288	alliance	1e-04	credible	1e-04
extraordinarily	0.0316	enhanced	1e-04	shield	0.0244
impressed	1e-04	achieve	1e-04		
		strengthen	1e-04		

Table 4.2.23: Negative emotion words that significantly contribute to the principal emotional signal

A Words	A p-values	LMNEG Words	LMNEG p-values	H4NEG Words	H4NEG p-values
concerns	1e-04	against	1e-04	make	1e-04
concern	7e-04	late	0.0053	get	1e-04
fears	1e-04	cut	4e-04	run	1e-04
worries	1e-04	losses	1e-04	recession	1e-04
warned	5e-04	decline	4e-04	default	1e-04
question	1e-04	concerns	1e-04	jobless	1e-04
uncertainty	1e-04	crisis	1e-04	try	1e-04
concerned	3e-04	weak	1e-04	volatility	1e-04
questions	1e-04	declined	1e-04	pass	1e-04
threat	1e-04	claims	1e-04	oversight	1e-04
avoid	1e-04	lost	1e-04	unexpectedly	1e-04
failure	1e-04	dropped	1e-04	exit	1e-04
fear	1e-04	unemployment	1e-04	risky	1e-04
worried	1e-04	problems	1e-04	mind	1e-04
rejected	1e-04	force	0.038	temporarily	1e-04
stress	4e-04	deficit	0.0011	ax	0.0016
worry	1e-04	recession	1e-04	spill	1e-04
uncertain	1e-04	sharply	2e-04	leak	5e-04
doubt	1e-04	default	1e-04	bankrupt	0.0089
threatened	1e-04	concern	3e-04	fragile	1e-04
fail	1e-04	bankruptcy	1e-04	disruption	0.0118
failing	1e-04	weakness	0.0117	unrest	0.0059
questioned	1e-04	failed	1e-04	bail	1e-04
threats	1e-04	fears	1e-04	aversion	1e-04
doubts	1e-04	slow	1e-04	suspend	0.0036
suspect	0.015	worries	1e-04	manipulation	1e-04
dangerous	1e-04	problem	1e-04		
feared	1e-04	hurt	1e-04		
fails	1e-04	difficult	1e-04		
distressed	1e-04	warned	0.0012		
danger	1e-04	declines	3e-04		
threatening	1e-04	question	2e-04		
threaten	1e-04	bad	1e-04		
anxiety	1e-04	outages	0.0143		
reject	0.0021	slowdown	6e-04		
panic	1e-04	concerned	2e-04		
warn	1e-04	poor	1e-04		
toxic	1e-04	slowing	4e-04		
hurdles	1e-04				
hurdle	1e-04				

Henceforth, the analysis will be carried out using the two purposely designed emotion lists, E and A, since they were both theoretically defined and appear to perform well empirically. Since the analysis of this section made use of the full available US Reuters database, it could be considered an act of overfitting for the purpose of testing *ex ante* predictions in an economic and financial setting if the RSS series was derived only using the words in Tables 4.2.22 and 4.2.23. Therefore, the full excitement and anxiety lists will be used. Alternatively, it would be possible to calibrate the lists empirically on a subset of the data and make ex ante predictions on the unseen dataset. However, this is left as a future research area.

4.2.5 Word-Frequency Adjustments

Thus far, the raw frequencies of individual words have been aggregated in order to produce the overall RSS series. However, the frequency distribution of words tend to be highly skewed, causing the overall signal to be largely dominated by the frequencies of a few number of words only (the most frequent words studied in the previous section). On the other hand, it is also the case that very infrequent words will yield less robust signals, due to the effect of a small sample, so care must be taken when increasing the contributions of very infrequent words.[7] From the lengthy analysis of the previous section it is now known, *ex post*, that the top words in the two lists constructed appear to consistently measure the respective emotional signals. However, *ex ante*, it would appear possible that usage of the most frequent words could deviate from the theoretical assumptions, due to noise or unexpected usage, which could potentially introduce noise to the emotional signal.

There are several potential ways to tackle some of these issues and improve on the methodology. The improvements will in general be more or less *ex post* in nature, if they rely on observations made on the full dataset. However, this is only a problem from the point of view of formal forecasting, not from the point of view of testing CNT as long as the selection of methodology is performed independently of the given tests.

One way to tackle the issue of a few words dominating the signal is to use a different word aggregation strategy which gives greater chance for low frequency words to contribute to the aggregate series. For example, one could normalise each word frequency series to have a common mean and common standard deviation and then average (equally or otherwise) the series to produce the final signal. By doing so, the long tail of low frequency words will likely introduce a lot of noise, but this noise might in fact cancel itself out assuming the words in the list are properly chosen. The question that arises is how to normalise the word frequency series for each word

[7] However, this effect will likely average out across all low frequency words.

without having future information creep into past data points. For example, if the standard z-score is computed using the full dataset for each word and the words are then averaged, a mean and standard deviation which would not have be available would have been used. Therefore, a rolling (using an 'expanding' window) mean and rolling standard deviation would be required. Both of which are computed by making use of all the data points available at the time.[8] Such normalised excitement and anxiety series are generated both by using all words in each list and, for comparison, also by only using the 50 most frequent words in each list. Table 4.2.24 displays the correlations between each variant. The normalised versions (in which all words are scaled, using rolling means and standard deviations, and then equally averaged to produce the final index) start in February 1996 since at least two data points are needed to compute the standard deviation.

Table 4.2.24: Correlations between different versions of the excitement and anxiety series; EXC and ANX represent the original series, Norm represents normalisation of all words and Norm 50 represents normalisation of the 50 most frequent words of the list

	EXC	ANX	EXC Norm	ANX Norm	EXC Norm 50	ANX Norm 50
EXC	1	-0.143	0.633	-0.106	0.789	-0.134
ANX		1	0.137	0.932	0.054	0.946
EXC Norm			1	0.205	0.828	0.136
ANX Norm				1	0.124	0.957
EXC Norm 50					1	0.089
ANX Norm 50						1

If the anxiety series is normalised according to the procedure described above, by computing rolling z-scores for each word frequency series (or only those of the 50 most frequent words) and then averaging the scaled series equally, the results remain correlated with the original series (for which the raw frequencies are simply aggregated) at 0.932 (ANX Norm) and 0.946 (ANX Norm 50). The corresponding excitement series are only correlated at 0.633 and 0.789 respectively, indicating that a larger proportion of excitement words might introduce noise that is biased (in the sense that the noise does not average out to the same extent). This difference should be expected given the PCA analysis from the previous section.

4.2.6 Wordlist Stability

A bootstrap resampling technique can be used to compute confidence intervals around the summary statistic. It is possible to sample both the words from the predefined

[8] In other words, the window of observations used to compute the current value is increasing over time. Hence, the initial values are computed using fewer data points than the latter values.

emotion lists and the documents in the collection to generate such confidence intervals. The summary statistic can be repeatedly computed across many such samples to derive a bootstrap distribution from which statements of confidence can be inferred. Words can be bootstrapped to assess the sensitivity of the results to the choice of words. This procedure yields wider confidence intervals at times when the word frequency distribution is more concentrated over a smaller number of words. Documents can be bootstrapped to assess the sensitivity of the measure to how the words are distributed across the document collection. This procedure yields wider confidence intervals at times when the words are less evenly distributed across the documents, for example if a smaller set of documents contain a large proportion of the found emotion words. Only the sensitivity of the measure to the choice of words will be considered in this section (as the RTRS document collection is very large).

From the analysis of this chapter, in terms of the sensitivity to the word lists themselves, it is now possible to be fairly confident that both the excitement and anxiety lists contain words which are good representatives of the same emotional signal (especially the anxiety list). However, if no normalisation is performed on the word frequency series of individual words before aggregating them to the final index, due to the nature of word frequency distributions (which tend to follow power laws as noted above), the exclusion of some of the most frequent words could potentially make a substantial difference. If normalisation is performed this problem is likely to be less substantial. It is possible to estimate to what extent this is a problem by drawing random samples with replacement from the wordlists to generate synthetic lists and recompute the RSS index with the various aggregation procedures to quantify this potential source of noise. This bootstrap procedure makes it possible to approximate the sampling distribution of possible RSS series to generate confidence intervals of the index with respect to the appearance of particular words in the list.

Bootstrap distributions of correlations for the above aggregation procedures are estimated, as well as with a raw aggregation of the 50 most frequent words, for the excitement and anxiety lists. The correlations are Pearson r scores computed between a particular bootstrap sample and the original series.

Figure 4.2.11 displays box plots of the four bootstrap distributions generated by sampling from the excitement dictionary (or only the 50 most frequent excitement words) with replacement and computing the excitement index either by aggregating the raw frequencies or by computing the rolling z-score for each word before aggregating. Figure 4.2.12 displays corresponding plots generated using the anxiety list. The main box starts at the first and ends at the third quartile of the corresponding distribution. The central line represents the median value. The 'whiskers' of the plot extend 1.5 times the interquartile range of the box. Anything

outside the whiskers are considered outliers and are plotted using small circles.

Once more, note the relative stability of the anxiety list compared to the excitement list. The scaled versions, in which each word gets an equal effective contribution to the final index, are much more robust than the raw versions. Finally, the versions including only the 50 most common words are seemingly not as robust as those containing all words.

4.3 Discussion

This chapter presented the basic RSS methodology, in particular how the preselected lexicons were constructed and how they are used to produce the RSS score and how articles are selected and aggregated into a time series of a given frequency. The second part of the chapter evaluates the methodology on several criteria. The evaluation section discusses the effects of different ways to scale the relative frequency counts of excitement and anxiety words, the potential effect of basic negation, as well as more non-trivial issues such as the evolution of the relative ranks and frequencies of the words in a given list and, most importantly, the extent to which the words in a list tend to capture the same emotional signal. A thorough empirical evaluation of six different emotion lexicons has been carried out on all Reuters articles published in New York and Washington over the period January 1996 through October 2014. Different methods to scale the emotion word frequencies to account for the amount of text analysed have been evaluated and it was found that scaling by the total number of characters or words seems sensible. The potential impact of negation on the RSS index was tested and it was found that, at least when the number of documents is as large as for the US Reuters database, it has little to no impact on the changes of the final index. The performance of different word lists is compared on the criteria of how well the most frequent words of each list capture a target emotion signal defined as the principal component of the correlation matrix of the individual word frequency series. It is found that the excitement and anxiety lists perform well (especially the anxiety list), and that there is room for improving the lists in terms of excluding words that act as outliers (in the sense of not contributing to the principal component, or even contributing negatively) and including words that correlate with the main signal. Using the excitement and anxiety lists, a different word aggregation strategy that gives equal weight to each word is explored, thus avoiding the strong dependence of the RSS metric on the most frequent words. It is showed, using a bootstrap test, that such an aggregation procedure yields more robust (with respect to word choice) results.

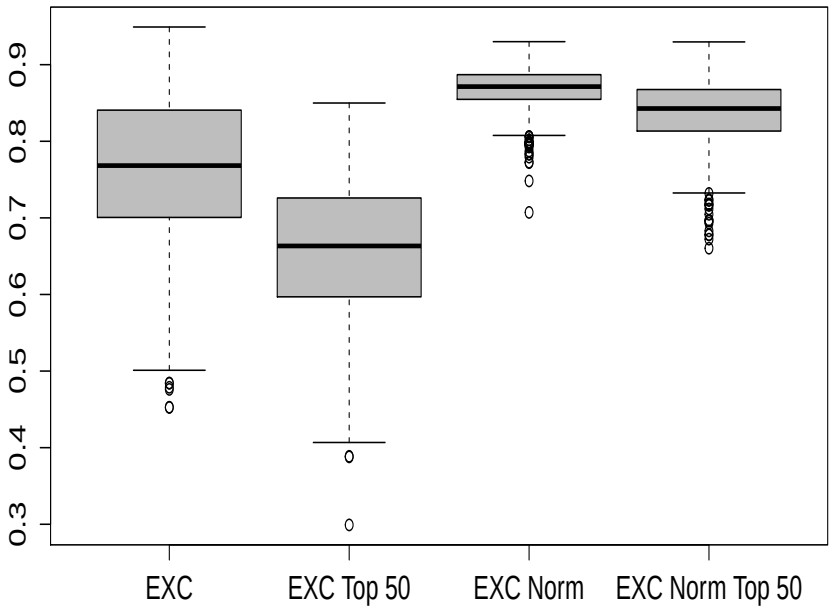

Figure 4.2.11: Bootstrap distributions of correlations between series generated from excitement word samples and the original US EXC series

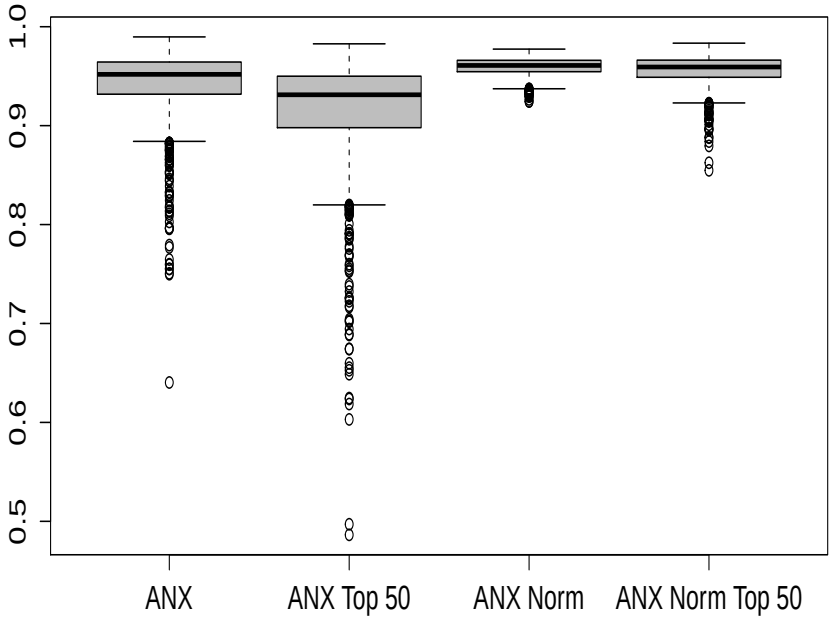

Figure 4.2.12: Bootstrap distributions of correlations between series generated from anxiety word samples and the original US ANX series

However, most of the analysis conducted relied on all information over the available period and might therefore not be suitable to apply in an *ex ante* test of predictability. The full excitement and anxiety word lists and the original methodology using the simple word frequency aggregation to produce the RSS series will therefore be used in the remaining chapters.

To conclude, the excitement and anxiety lexicons appear suitable to be used to make CNT operational as they are the two most consistent lists, and the anxiety list is also the list with the highest clarity score. The two lists and the RSS measure will now be used in the main experiments to follow this chapter.

5

Experiment 1: A Macro Measure of Relative Sentiment

The first hypothesis states that conviction and emotion drive investment, both in financial markets and in business, which ultimately fuels economic growth. Much investment activity must in some way manage uncertainty. CNT postulates that agents manage to act in the face of such uncertainty by gaining conviction through narratives. This hypothesis can be explored empirically using statistical techniques such as Granger-causality and regression analysis.

This chapter reports on the first experiment of the thesis using the derived relative sentiment shift methodology. The experiment explores the extent to which financial markets and the macroeconomy are driven by macro confidence in the form of aggregate relative sentiment. Before proceeding to test the hypothesis, indices of relative sentiment shifts will be compared to conventional measures of confidence and sentiment in the form of the Michigan Consumer Sentiment Index and the VIX.

5.1 Hypothesis

It is hypothesised that the relative balance between approach and avoidance (excitement and anxiety) determines action under uncertainty. As a consequence of this, the amount of investment made under uncertainty is expected to depend on RSS.

In other words, an increase in medium to long term financial and business investment (short term investment is not necessarily as dependent on conviction, although investment strategies themselves likely are - including quantitative and technical strategies) made under uncertainty should follow an increase in the relative difference between excitement and anxiety among economic actors. Similarly, a decrease in investment activity should follow from a decrease in aggregate relative emotion.

Given the complexity of financial markets and the economy as a whole, it is reasonable to assume that most medium to long term decisions have uncertain outcomes and should therefore be driven by conviction. This hypothesis is tested by constructing relative sentiment indices at the macro level and testing how well they predict or lead business investment (and as a consequence also GDP growth).

5.2 Methodology

The hypothesis will be tested using statistical techniques such as Granger-causality and regression analysis.

The tests of Granger non-causality will be based on the procedure of Toda and Yamamoto [109] as outlined in Chapter 2. Vector Autoregressive Models (VAR models) are specified for each pair of variables to be tested for causality. Wald tests are then used to assess the significance of lagged independent variables on explaining the dependent variable. The T-Y procedure is able to test for the presence of Granger-causality avoiding issues of 'spurious causality' [50] due to unit roots and/or co-integration between the two series. This is achieved by the addition of extra lags of each variable, corresponding to the maximum order of integration of the two series, before applying the Wald test on the lags of the independent variable. This procedure ensures that the test statistic has the assumed distribution under the null hypothesis even if the two series are co-integrated.

Once Granger-causality tests have been carried out, the nature of the relationship between the variables is explored through linear regression analysis.

Of course, more powerful tests of causality and the relationship between time series variables exist, for example information theoretic measures, which do not share some of the assumptions of linear regression (such as the assumption of linearity). However, if one is able to find a clear relationship using a very simple model the underlying hypothesis is more likely to be true. Every parameter introduced in a model will have its own associated estimation error potentially leading to a false conclusion of the main hypothesis.

Three different RSS indices will be considered, the US Reuters news based index constructed in Chapter 4, an index constructed from broker research reports and an

index constructed from an archive of financial market comments written by market experts at the Bank of England.

5.3 Results

Before proceeding to test the hypothesis, the relationship between various RSS indices created and conventional measures of market sentiment and consumer confidence will be explored.

5.3.1 Confidence

This section tests the potential of RSS to lead and predict more traditional measures of *confidence*. This includes the popular Michigan Consumer Sentiment Index (MCI) and the Chicago Board Options Exchange Market Volatility Index (VIX).[1]

Alongside the US RTRS RSS series derived from Reuters news articles published in New York and Washington, two further RSS measures will be introduced. One derived from an archive of broker circulars, or broker research reports, which will be referred to as BROKER, and one derived from an archive of market comments produced at the Bank of England, which will be referred to as MCDAILY. Details of the two databases can be found in Chapter 3. The series extracted for BROKER and MCDAILY are scaled by the total number of characters as opposed to the number of words. This was done to minimise the impact of numerical tables, which were present to a greater extent in these two sources than in Reuters news.

Table 5.3.1 shows the correlations between the three RSS measures, MCI and VIX. The correlations are computed over the available periods for the three RSS series (US RTRS; January 1996 through October 2014, BROKER; June 2010 through June 2013, and MCDAILY; January 2000 through July 2010).

To test for potential lead-lag relationships between the various series the T-Y procedure for Granger-causality testing is carried out as outlined in Chapter 2.

First, tests for the order of integration of each series are carried out over the period January 1996 through October 2014 (or a subset thereof given available data for BROKER, MCDAILY and the MCI). ADF and KPPS tests are performed with an increasing order of differencing until both tests suggest stationarity, the number of difference operations then performed is considered the series' order of integration.

[1]The VIX is a measure of implied volatility of S&P 500 index options. The index measures financial markets expectation of stock market volatility over the next 30 day period. The index is commonly referred to as the fear index. To ensure the index is comparable to the monthly RSS series a monthly average of the VIX will be considered as opposed to its value on any particular day.

Table 5.3.1: Correlations between US RTRS RSS, BROKER RSS, MCDAILY RSS, the Michigan Consumer Sentiment index and the VIX

	US RTRS	BROKER	MCDAILY	MCI	VIX
US RTRS	1	0.68***	0.59***	0.62***	-0.43***
BROKER		1	-	0.66***	-0.60***
MCDAILY			1	0.26***	-0.62***
MCI				1	-0.25***
VIX					1

Note: *p<0.1; **p<0.05; ***p<0.01

Table 5.3.2: Augmented Dickey-Fuller and Kwiatkowski-Phillips-Schmidt-Shin tests for stationarity of RSS series, the MCI and the VIX

Variable	ADF	p-value	KPSS	p-value
US RTRS Level	-3.272 (6)*	0.077	2.08 (3)***	<0.01
US RTRS Diff	-7.551 (6)***	<0.01	0.023 (3)	>0.1
BROKER Level	-2.677 (3)	0.31	0.466 (1)**	0.049
BROKER Diff	-3.975 (3)**	0.022	0.035 (1)	>0.1
MCDAILY Level	-1.678 (5)	0.71	0.70 (2)**	0.013
MCDAILY Diff	-6.66 (4)***	<0.01	0.046 (2)	>0.1
MCI Level	-2.299 (6)	0.45	3.948 (3)***	<0.01
MCI Diff	-6.577 (6)***	<0.01	0.063 (3)	>0.1
VIX Level	-3.112 (6)	0.11	0.282 (3)	>0.1
VIX Diff	-7.353 (6)***	<0.01	0.0298 (3)	>0.1

Note: *p<0.1; **p<0.05; ***p<0.01

From Table 5.3.2 it is possible to conclude that all series are integrated of order 1. In other words, the first order differences of each series is stationary whereas the levels are not. Therefore, an extra lag will be added to each VAR model when performing the Wald tests of non-Granger causality.

Secondly, the Akaike Information Criteria will be computed for each relevant VAR model, from 1 through 10 lags, to determine the optimal number of lags. Table 5.3.3 displays AIC scores for each model involving the MCI and Table 5.3.4 those models involving the VIX. The number of lags selected in each case is highlighted in bold.

Table 5.3.3: Akaike Information Criteria from 1 through 10 lags; VAR models with RSS and MCI (January 1996 through September 2014)

AIC(p)	US RTRS & MCI	BROKER & MCI	MCDAILY & MCI
1	1.959	**1.580**	-15.720
2	1.970	1.801	**-15.797**
3	**1.939**	1.903	-15.784
4	1.960	2.019	-15.734
5	1.980	2.142	-15.684
6	1.990	2.085	-15.671
7	2.021	1.642	-15.661
8	2.019	1.565	-15.665
9	2.037	1.538	-15.700
10	2.055	0.859	-15.734

Table 5.3.5 displays the test statistics for serial autocorrelation of residuals, using the multivariate Portmanteau- and Breusch-Godfrey tests, of the selected number of lags according to the AIC scores.

The stability of the selected models is checked by inspecting OLS-CUSUM plots, which all indicate model stability.

Finally, Table 5.3.6 reports the test statistics of the Wald tests for non-Granger causality in each direction for each VAR model.

There is evidence of Granger causality from the US RTRS series to the Michigan Consumer Sentiment index, but not vice versa. There is evidence of causality from the BROKER series to the Michigan Consumer Sentiment index but not vice versa. Finally, there is also weak evidence of Granger causality from MCDAILY to the VIX

Table 5.3.4: Akaike Information Criteria from 1 through 10 lags; VAR models with RSS and VIX (January 1996 through October 2014)

AIC(p)	US RTRS & VIX	BROKER & VIX	MCDAILY & VIX
1	1.817	2.303	-15.855
2	1.711	**2.380**	**-16.041**
3	1.675	2.575	-16.027
4	**1.680**	2.646	-15.997
5	1.689	2.341	-15.954
6	1.676	2.525	-15.924
7	1.705	2.386	-15.882
8	1.718	2.309	-15.852
9	1.749	2.428	-15.825
10	1.780	2.110	-15.796

Table 5.3.5: Portmanteau and Breusch-Godfrey tests for serial correlation of residuals; VAR models with RSS, VIX and MCI (January 1996 through October 2014)

VAR model	Portmanteau	d.f.	Breusch-Godfrey	d.f.
US RTRS/MCI	47.045	52	18.6741	20
US RTRS/VIX	51.471	52	26.109	20
BROKER/MCI	41.681	60	15.596	20
BROKER/VIX	40.522	56	25.613	20
MCDAILY/MCI	60.799	56	23.738	20
MCDAILY/VIX	52.535	56	22.577	20

Note: *p<0.1; **p<0.05; ***p<0.01

Table 5.3.6: Wald test statistics of Granger-non-causality; VAR models with RSS, VIX and MCI (January 1996 through October 2014)

Direction	Chi-Sq	d.f.	p-value
US RTRS → MCI	8.2	3	0.043**
MCI → US RTRS	0.69	3	0.88
US RTRS → VIX	3.0	3	0.4
VIX → US RTRS	2.2	3	0.53
BROKER → MCI	44.9	1	2.1e-11***
MCI → BROKER	0.005	1	0.94
BROKER → VIX	3.4	2	0.18
VIX → BROKER	3.7	2	0.16
MCDAILY → MCI	1.4	2	0.5
MCI → MCDAILY	0.11	2	0.95
MCDAILY → VIX	4.8	2	0.092*
VIX → MCDAILY	1.9	2	0.39
Note:	*p<0.1; **p<0.05; ***p<0.01		

but not vice versa.

More formal *ex ante* forecasts of the MCI using the BROKER index can be found in [78]. A simple linear regression model, with the change in the Michigan Consumer Sentiment index from the previous final publication (at time $t-1$) to the current preliminary publication (at time t) as the dependent variable and the change in RSS within the BROKER archive in the preceding period ($RSS_{t-1} - RSS_{t-2}$) as the independent variable, is estimated using data up to time t. The parameters of the estimated model are used to forecast the change from the final MCI at time t to the preliminary estimate in time $t+1$. In other words, only information available at time t is used to make a prediction of information available in time $t+1$. The procedure is then moved forward one period and repeated. Thus, 15 such predictions are constructed and the forecasting accuracy is reported in Table 5.3.7. In comparison, the accuracy of consensus forecasts made by economists and published in Reuters is reported in Table 5.3.8. The various forecasts and the true value of the index are plotted in Figure 5.3.1.

Forecasts made using the BROKER index and linear regression models achieve an

Table 5.3.7: Forecasting accuracy of MCI using BROKER; regressing the true values on the forecasts

	Dependent variable:
	TRUE VALUE
PREDICTION	1.219***
	(0.323)
Constant	0.347
	(0.852)
Adjusted R^2	0.486
Residual Std. Error	3.081 (df = 13)
F Statistic	14.245*** (df = 1; 13)
Note:	*p<0.1; **p<0.05; ***p<0.01

Table 5.3.8: Forecasting accuracy of MCI CONSENSUS forecasts; regressing the true values on the forecasts

	Dependent variable:
	TRUE VALUE
PREDICTION	1.972
	(1.178)
Constant	−1.293
	(1.084)
Adjusted R^2	0.114
Residual Std. Error	4.045 (df = 13)
F Statistic	2.804 (df = 1; 13)
Note:	*p<0.1; **p<0.05; ***p<0.01

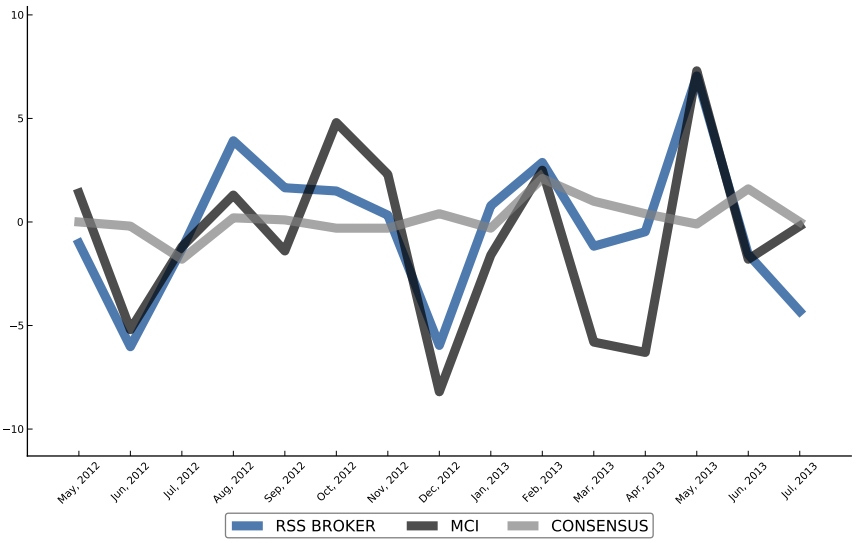

Figure 5.3.1: Forecasts of the change in the preliminary estimate of Michigan Consumer Sentiment index since previous final estimate; CONSENSUS vs. BROKER RSS

adjusted R squared of 0.49, compared with 0.11 for the consensus economist forecasts. In other words, forecasts using the RSS index explain more than 4 times as much of the variation in the true changes in the Michigan index than the consensus economist forecasts. As can be seen from Table 5.3.8, the consensus forecasts cannot be distinguished from noise (the p-value on the coefficient of the forecasts is not below 0.1). The predictions made using the BROKER index are unbiased, since the coefficient on the predictions are not significantly different from one and the constant term is not significantly different from zero.

The next section presents the tests of the main hypothesis, the effect of macro relative sentiment on economic growth.

5.3.2 The Macroeconomy

To what extent can economic growth be explained and predicted by the US RTRS RSS index.[2] Reuters news archive can be filtered for those articles published by Reuters in various regional offices, such as New York and Washington. Since the goal is to measure the aggregate emotion of the agents in a particular region, the assumption is made that journalists in a particular region will use language that reflects their emotional state and indirectly also that of the agents in their surroundings. The index constructed for the US by analysing articles published in the

[2]The econometrics in this section is taken from [115] and is analysis carried out together with Paul Ormerod.

region is therefore expected to reflect the emotional state of the local agents and therefore influence the US economy. In other words, it is believed that the US RTRS index captures the confidence of the US economy and that this confidence will feed into new investments and therefore also GDP growth. Thus, a claim about causality is made as opposed to a claim about correlation (or that RSS is simply a more timely indicator of GDP growth and therefore correlates with or appear causal with GDP), or indeed a claim that RSS simply reflects fundamental information. It is therefore necessary to deal with a potential problem of endogeneity. In other words, to deal with the question of whether or not RSS simply measures the effect of relevant fundamental information, present in the news, on GDP growth. It can be argued that emotions have no impact on economic growth and that they simply are responses to *informational shocks*. Such shocks are typically considered exogenous to the economic system by standard economic theory and that they cannot be modelled or predicted. This problem is dealt with in two different ways. First of all, by considering the effect of RSS on growth while controlling for various asset price variables thought to capture relevant fundamental information. Secondly, by constructing alternative 'fundamental series' using common economic terms such as 'crisis', 'recession', 'bankruptcy', etc. instead of via the emotion lexicon, and exploring a potential causal relationship between this variable and RSS. The impact of relative sentiment (and thereby CNT) on economic growth can be seen much more clearly after disentangling the underlying emotion and other potentially relevant informational content.

GDP Growth

A key challenge to macroeconomists, central banks and policy-makers is to forecast economic growth (gross domestic product, or GDP). This task is made especially hard due to the various stages of revisions made to GDP data, as well as the inherent degree of randomness of GDP [80]. This section shows how CNT can improve the understanding of the evolution of GDP growth and how RSS can be used as a new source of information to be included into economic growth forecasting models.

Two major components of GDP are personal consumption expenditures and gross private domestic investment. These components can be further divided into subcomponents, as seen in the below list.

Gross Domestic Product and its component parts

Personal consumption expenditures

Goods

Durable goods

Nondurable goods

Services

Gross private domestic investment

Fixed investment

Change in private inventories

Net exports of goods and services

Government consumption expenditures and gross investment

The component most expected to be caused by shifts in relative sentiment is fixed investment by firms. Any payoffs of such investments tend to be highly uncertain in nature and such investment decisions are therefore likely to require emotional conviction. In particular, the relationship between Gross Private Domestic Investment made by firms and the total level of GDP i.e., $\frac{GrossPrivateDomesticInvestment}{GDP}$ is expected to be significant. An economy that spends more on investment and less on other components of GDP is generally interpreted as engaging in a more dynamic, aggressive form of economic development. Put differently, if firms decide to invest more given a constant level of GDP they likely expect higher growth, and as such demand for products and services, in the future. Conversely, if firms decide to invest less given a constant level of GDP they expect lower growth in the future.

In terms of personal consumption expenditures, it might be expected that spending on durable goods could be determined by relative sentiment to some extent whereas other forms of personal consumption spending should be less influenced by it. Government consumption expenditures and gross investment would not be expected to follow relative sentiment at all.

Some evidence is provided that this problem might successfully be approached using Granger causality tests and simple regression with the US RTRS RSS index constructed above. Quarterly relative sentiment series from 1996Q1 through 2014Q3 are created by aggregating the documents on a quarterly basis before computing the RSS statistic for each such quarter. The series is plotted together with GDP growth in Figure 5.3.2. The two series clearly appear to be correlated, and indeed the correlation between the quarterly RSS series and GDP growth (DLGDP - computed by taking

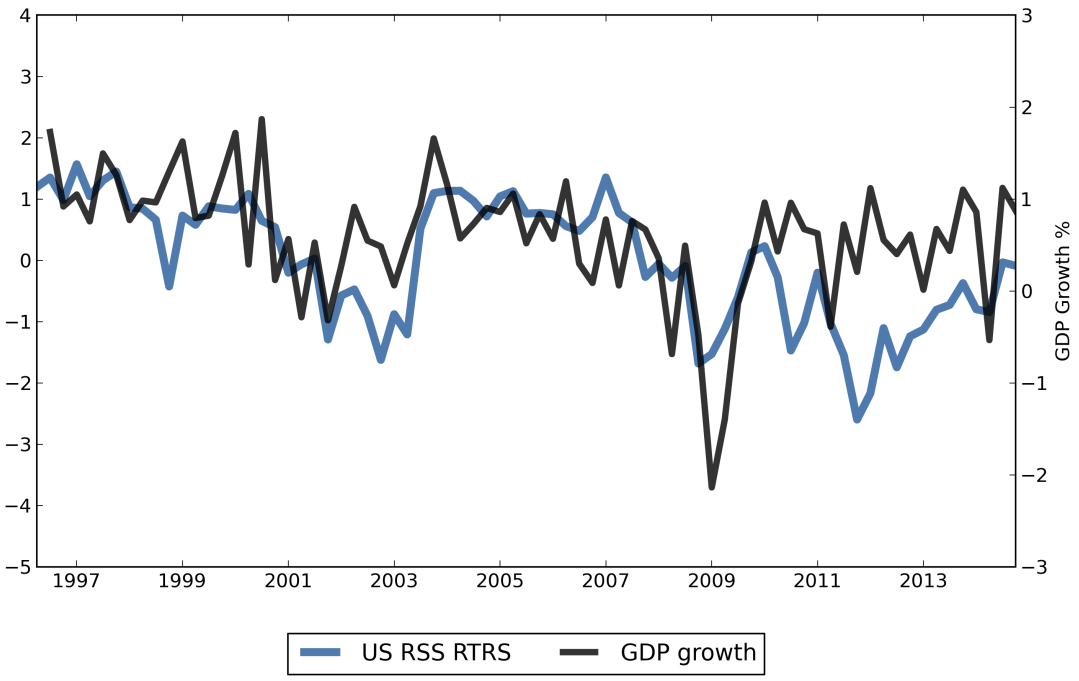

Figure 5.3.2: US RSS RTRS and US GDP Growth; 1996Q1 through 2014Q3

the first order difference of log levels of GDP) over the period 1996Q1 through 2014Q3 is 0.48, and the correlation between RSS and $\frac{GrossPrivateDomesticInvestment}{GDP}$ is 0.52 and RSS and changes in consumer expenditure (DLCE - computed by taking the first order difference of log levels of consumer expenditure) is 0.54.

Tests of Granger non-causality are carried out between RSS and quarterly DLGDP, DLCE and $\frac{GrossPrivateDomesticInvestment}{GDP}$ (GPDIGDP), in the United States.

Table 7.3.6 presents the results of step 1 of the T-Y procedure involving US RSS, US DLGDP, US GPDIGDP and US DLCE.

The ADF tests are all conducted with 4 lags and the KPSS tests with truncation lag parameter equal to one (as indicated by the figures in brackets).

All variables need to be differenced once before both the ADF and the KPSS test indicate that the series are stationary, with the exception of GPDIGDP. However, from a theoretical perspective, it is known that this variable is bounded between 0 and 1 (since it is a percentage of total GDP) and that it therefore must have finite mean and variance. It is therefore unlikely that this variable truly is of greater order of integration than 1, and it will therefore be treated as such. In other words, all series are believed to be integrated of order 1, i.e. are $I(1)$. Therefore, in all tests

Table 5.3.9: Augmented Dickey-Fuller and Kwiatkowski-Phillips-Schmidt-Shin tests for stationarity of RSS series, DLGDP, GPDIGDP and DLCE (1996Q1 through 2014Q3)

Variable	ADF	p-value	KPSS	p-value
RSS Level	-2.25(4)	0.47	1.48(1)***	<0.01
RSS Diff	-4.20(4)***	<0.01	0.04(1)	>0.1
DLGDP Level	-2.78(4)	0.25	0.78(1)***	<0.01
DLGDP Diff	-5.04(4)***	<0.01	0.03(1)	>0.1
GPDIGDP Level	-2.365(4)	0.43	1.33(1)***	<0.01
GPDIGDP Diff	-3.05(4)	0.15	0.22(1)	>0.1
DLCE Level	-2.65(4)	0.31	1.40(1)***	<0.01
DLCE Diff	-4.13(4)***	0.01	0.03(1)	>0.1

Note: *p<0.1; **p<0.05; ***p<0.01

Table 5.3.10: Akaike Information Criteria from 1 through 10 lags; VAR models with RSS and DLGDP, GPDIGDP and DLCE (1996Q1 through 2014Q3)

AIC(p)	RSS & DLGDP	RSS & GPDIGDP	RSS & DLCE
1	**-11.166**	-11.927	-12.133
2	-11.096	**-12.019**	-12.116
3	-10.988	-11.983	**-12.171**
4	-10.900	-11.890	-12.072
5	-10.819	-11.820	-12.045
6	-10.791	-11.784	-12.082
7	-10.768	-11.815	-12.077
8	-10.654	-11.762	-12.030
9	-10.676	-11.686	-11.942
10	-10.587	-11.611	-11.859

involving a combination of the above variables, m in step 1 above is chosen to be one.

The AIC values for $VAR(p)$ models of the relevant pairs of variables, with the number of lags p varying from 1 through 10 (it would not be expected for the model to have more than 10 lags of quarterly data) are shown in Table 5.3.10. For the model with DLGDP AIC is found to be minimised when $p = 1$. For the model with GPDIGDP AIC is found to be minimised when $p = 2$. For the model with DLCE AIC is found to be minimised when $p = 3$. The corresponding $VAR(p)$ models for all relevant pairs of variables are therefore fitted as such. In all cases, it is not possible to reject absence of serial correlations of residuals (as seen in Table 5.3.11) and OLS-CUSUM plots show parameter stability.

Table 5.3.11: Portmanteau and Breusch-Godfrey tests for serial correlation of residuals; VAR models with RSS, DLGDP, GPDIGDP and DLCE (1996Q1 through 2014Q3)

VAR model	Portmanteau	d.f.	Breusch-Godfrey	d.f.
RSS/DLGDP	44.72	60	14.90	20
RSS/GPDIGDP	40.56	56	20.72	20
RSS/DLCE	39.09	52	21.22	20

Table 5.3.12: Wald test statistics of Granger-non-causality; VAR models with RSS, DLGDP, GPDIGDP and DLCE (1996Q1 through 2014Q3)

Direction	Chi-Sq	d.f.	p-value
RSS $\rightarrow$ DLGDP	1.1	1	0.28
DLGDP $\rightarrow$ RSS	0.002	1	0.96
RSS $\rightarrow$ GPDIGDP	6.9	2	0.032**
GPDIGDP $\rightarrow$ RSS	1.6	2	0.45
RSS $\rightarrow$ DLCE	0.32	3	0.96
DLCE $\rightarrow$ RSS	6.3	3	0.10
Note:	*p<0.1; **p<0.05; ***p<0.01		

Table 5.3.12 shows the test statistics of the relevant Wald tests, their degrees of freedom and the corresponding p-values. The null of Granger non-causality is rejected in one out of the six tests performed. In particular, RSS is Granger-causing

GPDIGDP, but not the other variables. Neither of the three GDP variables are found to Granger-cause RSS. These findings support the hypothesis that relative sentiment shifts within Reuters news archive, eventually feed into economic growth. In particular, there is evidence that changes in the degree of confidence can be used to predict changes in gross private domestic investments made by firms as a proportion of the total level of GDP. There appears to be no clear relationship between confidence as measured and consumer spending. This is in line with expectations from CNT. In other words, confidence appears to have a causal effect on economic growth, which appears to be through private investments made by firms and not consumer spending.

A question that is often raised regarding the use of sentiment indicators to explain economic variables is to what extent the sentiment indicators simply reflect information contained in conventional forward looking measures, such as asset prices. A more generic multivariate linear regression equation for US GDP growth is therefore constructed by including a range of common asset price variables, as well as RSS. The following variables are considered.

- The Michigan Consumer Sentiment Index (MCI)
- Changes in stock prices, measured by the Standard and Poors 500 (DSP)
- The yield on 10 year US government bonds (BOND)
- The yield on 3 month US Treasury bills (TB)
- The TED spread (TED)
- The VIX, the Chicago Board Options Exchange Market Volatility Index

A completely generic multivariate equation including up to 3 lags of each variable listed above is estimated, in the following form,

$$\begin{aligned} DLGDP_t = \alpha + \sum_{i=1}^{3} [\beta_i MCI_{t-i} + \gamma_i DSP_{t-i} + \delta_i BOND_{t-1} + \theta_i TB_{t-i} + \\ \lambda_i TED_{t-i} + \phi_i VIX_{t-i} + \psi_i RSS_{t-i} + \omega_i DLGDP_{t-i}] + \epsilon_t \end{aligned} \tag{5.1}$$

One by one, the variable with the highest p-value is excluded and the regression re-estimated until only significant variables remain. If, at any stage in the process, two different lags of the same variable have equally sized but opposite sign on their estimated coefficients, the two variables are replaced by their difference and the

Table 5.3.13: US GDP equation with US RTRS RSS and asset-price variables; 1996Q4 through 2014Q3

	Dependent variable:
	$DLGDP_t$
$RSS_{t-1} - RSS_{t-2}$	0.002^{**}
	(0.001)
RSS_{t-3}	0.002^{**}
	(0.001)
$TED_{t-1} - TED_{t-2}$	-0.010^{***}
	(0.003)
TED_{t-3}	-0.004^{**}
	(0.002)
DSP_{t-2}	0.00004^{***}
	(0.00001)
Constant	0.007^{***}
	(0.001)
Observations	71
R^2	0.395
Adjusted R^2	0.348
Residual Std. Error	0.005 (df = 65)
F Statistic	8.475^{***} (df = 5; 65)
Note:	*p<0.1; **p<0.05; ***p<0.01

process continued. The result is the generic equation for US GDP growth reported in Table 5.3.13.

The equation is well-specified as seen from the various test statistics: Breusch-Godfrey test for serial correlation of order up to 4 (1.358; p-value 0.852), Breusch-Pagan (8.605; p-value 0.126), Durbin-Watson (1.769; p-value 0.120), RESET (0.876; p-value 0.421), Kolmogorov-Smirnov test for normality of residuals (0.096; p-value 0.498).

Does the US RTRS series simply reflect fundamental news? From the analysis of the principal components of the various emotion lexicons it became apparent that a large proportion of the most frequent words in the excitement and anxiety lexicons tend to move similarly and that the principal components themselves likely capture a common latent emotional variable rather than context specific informational shocks (as such shocks would likely make the words in a given list have uncorrelated movements). As a further indication that this is the case, a new time series is created by counting the frequency of a collection of economic terms that might also be good predictors of the business cycle. It is shown in other work by the author [115] that the RSS series Granger-causes this variable representing fundamental economic news, and not vice versa. It is also shown in the same work that the 'economic' index does not Granger-cause any of the GDP series. Thus, it is highly unlikely that the RSS index simply reflects fundamental news.

5.4 Discussion

CNT postulates that agents are able to act despite being faced with uncertainty (regarding the outcome of the given action) by creating narratives that yield conviction. The relative balance between excitement about gain and anxiety about loss is therefore believed to be a key determinant for action under uncertainty. The theory hypothesises that an increase in excitement relative to anxiety for the economy as a whole, which can be seen as increased economic confidence, translates into more economic activity in areas faced with high uncertainty. Much economic activity, in particular investment by firms, is conducted under real uncertainty where no probability distribution can be derived for the eventual outcome. For example, if a manufacturing company should invest in new machinery or a company is deciding on whether or not to hire more staff. By analysing news articles it is possible to investigate the hypothesis that the aggregate balance between excitement and anxiety leads measures of economic growth.

The relationship between various RSS measures derived from different databases and more traditional indicators of sentiment, the Michigan Consumer Sentiment index

(MCI) and the VIX, is explored. The analysis shows that the US RTRS series and the BROKER series both clearly Granger-cause the MCI. Furthermore, the RSS series extracted from market comments written at the Bank of England weakly Granger-causes the VIX.

Tests are then carried out to determine whether or not the RSS series can be used to forecast GDP growth, as hypothesised by CNT. The hypothesis is explored using Granger-causality tests and multivariable regression. It is found that the index constructed can be used in formal forecasting equations of US GDP growth, even after controlling for other common forward-looking asset-price variables. Potential issues regarding endogeneity, i.e., that fundamental information about the economy might be what is actually measured by RSS as supposed to confidence, are explored by comparing the RSS index with an index constructed using economic terms. It is found that the RSS index leads this 'fundamental' index. The results greatly support a central concept of CNT, that a key variable effecting the degree of investment activity among firms is the degree of conviction (or confidence) among economic actors.

The next chapter presents a couple of case studies that illustrate the concepts of *phantastic objects*, *groupfeel* and *divided states* on the level of individual narratives, in contrast with the macro analysis conducted in this chapter.

6

Experiment 2: A Micro Measure of Relative Sentiment

This chapter deals with three important aspects of CNT that can come to characterise some narratives.[1] In Chapter 2.1 the concepts of *phantastic objects*, *divided states* and *groupfeel* were discussed. The first two of these concepts refer to particular narratives that come to deviate substantially from reality due to an excess of relative excitement and/or a lack of relative anxiety. The objects of such narratives are referred to as phantastic objects and the state of mind in which reality is not adequately represented, such that positive and negative aspects are not properly balanced, is referred to as a divided state. Such narratives could be thought of as 'toxic' narratives, in the sense that decisions based on them will likely have negative outcomes and negative effects on the individual's mental and physical well-being, and in the sense that they tend to spread socially due to groupthink, groupfeel and social conventions. This chapter presents case studies of narratives that are expected, ex post, to have some of these characteristics. The aim is to find out if the RSS methodology can be applied to recognise them. This is a necessary first step to be able to apply the methodology in future research to attempt to detect such toxic narratives ex ante.

First, an analysis of Reuters News archive focusing on narratives about Fannie Mae

[1]The material in this section is published in [114]. I would like to acknowledge the contributions of Professor David Tuckett and Dr Rob Smith in designing the experiment.

is presented, representing the narrative of new housing finance that played a significant role in the 2008 financial crisis. The second case study concerns narratives about two business ideas at Enron Corporation the years around the time of the company's bankruptcy, presented via an analysis of a subset of their email database. The case study on Fannie Mae shows very clearly how the market entered into a divided state by becoming substantially less anxious about stories involving Fannie Mae even though available facts, in the form of declining housing prices, should have made investors more anxious. The Enron database makes it possible to divide the network of Enron staff, by who where sending emails to whom, into different social groups and show how for these two particular narratives one of the most important social groups had a significantly different emotional pattern from the rest of the network. The stark difference between social groups clearly showed the outcome of groupthink and groupfeel behaviour. It is shown that patterns consistent with phantastic object narratives can be detected and tracked as they develop and spread through networks to lead to a disconnect between narrative and underlying reality.

6.1 Phantastic Object 1: Fannie Mae

Fannie Mae is the oldest US entity offering large-scale, more or less government-backed, housing finance. It created mortgage-backed securities to become one of the largest mortgage financiers selling the securities to major banks in the years leading up to 2007. Its loans were not the typical subprime loans securitised by the large issuers and it did not touch some of the inferior credit claims sold by others [98]. Founded by the government, Fannie Mae was later privatised and has been a publicly-traded company since 1968, albeit that its exact relation to government was a matter for conjecture. When credit markets froze in mid-2007 Fannie Mae had difficulties and eventually collapsed into bankruptcy and was taken over by the Federal Government in the late summer of 2008. As a very large issuer of housing finance, it would be expected (on the grounds of 'rational expectations') that stories about Fannie Mae in the period before and after the financial crisis contain such excitement or doubt about what was happening in the securitised mortgage market or the housing market as real events would stimulate.

6.1.1 Hypothesis

Given the significant role of new housing finance leading up to the financial crisis of 2008, and how it lured the market into believing in a new growth and financial regime, CNT would expect stories about Fannie Mae, an entity deeply involved in this business, to exhibit a diminishing amount of anxiety relative to excitement leading up

to the crisis and that this trend would go on for much longer than what could be accounted for by 'rational expectations'. The hypothesis is that given available information about the state of the housing market, relative sentiment about Fannie Mae would show diminishing anxiety beyond a point that can be considered 'rational', but one that supports decision-making as postulated by CNT.

6.1.2 Methodology

The analytical framework developed for the RSS methodology will be used and focused on particular narratives. Simple pattern matching is relied on to filter the full English Reuters archive for those articles that mention Fannie Mae.

The filtering methodology of Definition 1 and 2 will again be applied. The pattern used is 'Fannie\s+Mae', which matches the word pair with any number of white spaces, or new-line characters, in-between. Furthermore, the word search proximity is defined to be the same sentence containing the pattern. In other words, the method only counts excitement and anxiety words within the same sentence (excluding any sentence found to be longer than 500 characters, which would indicate that the text would either not have been split properly or that the sentence might not only refer to Fannie Mae) as the pattern. In this case, the word frequency difference is scaled by the total number of sentences that matches the pattern.[2]

This will result in a relative sentiment shift series focused on Fannie Mae. This RSS series can be compared to an index reflecting the state of the actual US housing market, the Case-Shiller house price index. Through this comparison, a significant positive trend in RSS surrounding Fannie Mae over the relevant period leading up to the crisis is found, and that this trend continued substantially beyond the point when housing prices started to fall. Thus showing that the emotion within the narrative about Fannie Mae did not 'rationally' reflect this change in fundamental information about the housing market, indicating a clear divided state.

The significance of a trend in RSS is measured using a bootstrap procedure. To determine whether a sentiment trend is significant (relative to what could be observed in a random sample of articles) samples without replacement are drawn from a comparative universe of articles. An empirical distribution of the statistic of interest can easily be generated from these samples.

Formally, by applying Definitions 1 and 2, an empirical distribution of possible values given a number of Monte Carlo iterations α is defined as follows.

Definition 3. *Empirical distribution*

[2]This study was performed and published in Social Networks [114] prior to the work analysing the effects of different scaling strategies more formally on US Reuters articles.

Fix a number of Monte Carlo iterations α. Let T represent the entire collection of texts. Let $C_1, \ldots, C_n$ and $C'_1, \ldots, C'_m$ be some database fields. Let $f(t)$ be a function on collections of texts, $t \subseteq T[C_1, \ldots, C_n]$ with domain T given by any conditional set of texts. Let $k \in \mathbb{N}$ be any integer.

Define an empirical distribution *conditioned on $f, C_1, \ldots, C_n, C'_1, \ldots, C'_m$ and k up to approximation level α by the set*

$$\begin{aligned} & MC(f, T[C_1, \ldots, C_n], k, \alpha) \\ = \{ & f(sample(k, T[C_1, \ldots, C_n])) \\ | \ & sample \text{ is repeatedly applied } \alpha \text{ times} \} \end{aligned} \tag{6.1}$$

where $MC(f, T, k, \alpha)$ will be used as the notation for performing α Monte Carlo iterations using function f applied to the random sample of k items from set T, and sample(k, T) is a function which randomly samples k items from set T.

The definition is straight forwardly extended to functions with more arguments. With this empirical distribution one can apply significance testing.

Definition 4. *Monte Carlo based significance testing*

With notation as above, let P be the probability that a randomly chosen collection of k texts from T would generate a number as extreme as $f(X)$ for particular collection of texts X under the function f. Let $MC(f, T, k, \alpha)$ be a Monte Carlo generated empirical distribution for a given approximation level $\alpha \in \mathbb{N}$. Thus, P may be approximated by

$$P = \frac{q+1}{\alpha+1},$$

where q is the number of random samples getting a test statistic at least as extreme as $f(X)$.

The approximation level α determines the degree of accuracy of the empirical distribution. The only way to avoid any approximation is to generate all possible values of the measure. This number depends on the sample size k, the size of the entire collection under consideration T, and whether or not f is an injective function. Without this knowledge about f it would be necessary to apply f on $\binom{|T|}{k} = \frac{|T|!}{(|T|-k)!k!}$ subsets of T to get an exact distribution. As done in Chapter 4 when estimating the p-values of principal component contributions, a biased p-value is used rather than the unbiased $\frac{q}{\alpha}$ since random samples of the full sample distribution are considered.

6.1.3 Results

Figure 6.1.1 displays the RSS series of Fannie Mae (as well as the corresponding anxiety and excitement series) smoothed with exponential smoothing parameter set at 0.1. In other words, the smoothed value assigned to period t, y_t, is the weighted sum of the previous smoothed value y_{t-1} and the raw value assigned to period t, x_t, with weights 0.9 and 0.1 respectively.

Figure 6.1.2 shows z-scores of RSS in sentences that mentions Fannie Mae, the share price of Fannie Mae and the Case-Shiller 10 city seasonally adjusted house price index. The visual trends are clear:

1. For two years from early 2005 until mid-2007, stories about Fannie Mae were markedly different. The balance of excitement and anxiety within them shifted as they became increasingly free of anxiety words, relative to excitement words. A trend in relative sentiment of this kind was unusual and clearly indicates that a potential divided state had developed as hypothesised.

2. At the beginning of the 4th quarter in 2008, precisely as the financial crisis broke, there is a massive reversal and a catastrophic increase in anxiety and fall in the share price of Fannie Mae as well as a collapse in the housing price index. The relative sentiment shift is that predicted for the collapse of a divided state.

3. Over the period early 2005 until mid-2007 Fannie Mae's share price rose. In general the share price and relative sentiment index seem well correlated over this period. Investor decisions about the price of Fannie Mae in this key period, therefore, seem to have been determined by sentiment.

4. When the relative sentiment index is plotted against the Case-Shiller index of housing prices, something particularly interesting is observed. Whereas there is usually a positive correlation between the house price index and relative sentiment in the stories about Fannie Mae and Fannie Mae's share price, during the vital period 2005-2007 this correlation appears to have become negative. The explanation is that for about 14 months prior to the outbreak of the financial crisis, when house prices levelled off or fell (with what on logical grounds would be serious implications for Fannie Mae and mortgage finance), relative sentiment continued to become more excited, actually showing a continuing trend of decreasing anxiety. In other words, it is appears to be sentiment, rather than the relevant facts about house prices, which has captured investment opinion.

These visual trends are now confirmed statistically.

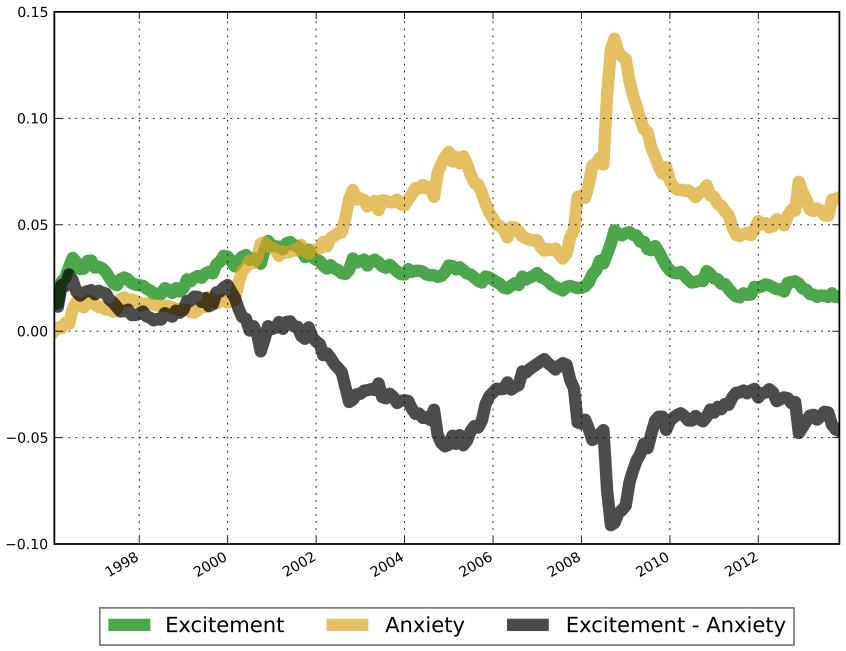

Figure 6.1.1: Fannie Mae RSS, Excitement and Anxiety series

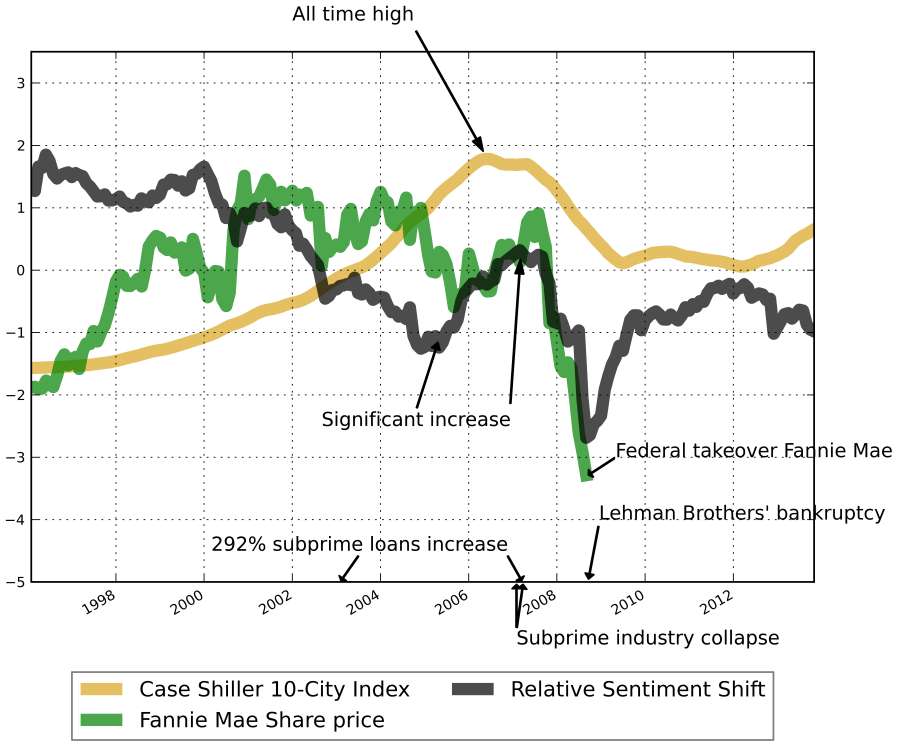

Figure 6.1.2: Fannie Mae RSS, Case-Shiller 10-City seasonally adjusted house price index and Fannie Mae share price

First, a test is performed of the statistical significance of the increase in relative sentiment for Fannie Mae (relative to other stories) in the period leading up to the crisis. It is hypothesised by CNT that relative sentiment about Fannie Mae in the period leading up to the financial crisis would exhibit diminishing anxiety relative to excitement. The period from April 2005 to February 2007 (inclusive) is tested, which is a conservative estimate of the pre-crisis exuberance (the relevant period for testing the hypothesis stated earlier in the chapter). Note yet again that both the object, Fannie Mae, and the period for which a divided state was expected to have emerged were selected with ex-post knowledge of the events. This experiment is performed as an initial test that the methodology is able to detect the patterns hypothesised by CNT. A future goal, although highly ambitious, is to apply the methodology to warn about emerging phantastic objects ex ante.[3] To quantify the increase in sentiment, the difference between the value at the end and the value at the start of the period is measured (recall that the series is smoothed which makes this statistic account for the values in-between the periods as well). A bootstrap sampling strategy will be carried out to estimate the significance of the increase. The sample space is set to articles published in the same period, thus taking into consideration any sentiment bias present in the given time-period. This is a critical step in order to distinguish between the target topics and general economic activity occurring simultaneously, which can be considered relatively exuberant.

To test the hypothesis sentences are randomly sampled from two sub-collections of all Reuters news articles. One collection contains all sentences from articles published in the relevant period and one collection contains all sentences from articles published in the relevant period *that* match the pattern for Fannie Mae. The same analysis as performed on the relevant Fannie Mae texts is performed on the random samples, thus generating two empirical distributions of the relevant statistic (a change in sentiment over the given period). Note that whatever collection is chosen as the sample space, it will inevitably condition the resulting distribution. However, by considering several collections one can be more confident about whether or not the result is found to be significant. It is important to sample the same number of texts (i.e., sentences) as those used in the target analysis to avoid any sample-size bias.

To test for significance of the relative sentiment increase define the function f in Definitions 3 and 4 to take several arguments, one for each data point in the relative sentiment series leading up to the end period. The function computes the relative sentiment of each argument and constructs a moving average time-series before

[3]It is interesting to note that relative sentiment and share price appeared to have become more correlated during this pre-crisis period (from about mid-2005) than before. This behavioural pattern could potentially be a candidate for a financial 'bubble' indicator to be explored in future research, as suggested by an anonymous reviewer.

computing the increase between the start and end of the considered period. With this definition, a random relative sentiment series is constructed in exactly the same way as the Fannie Mae series.

Figure 6.1.3 depicts the results of the two generated distributions, based on 10000 samples from each of the two different collections of texts. One in which all sentences within all English articles published in the same period as the target period are considered (on the right) and one in which only the sentences within the articles containing the Fannie Mae sentences are considered. Note that whatever distribution is considered the same conclusion is arrived at, namely that the shift in Fannie Mae relative sentiment for the target period was highly significant. Note also that the distribution constructed from the Fannie Mae articles is biased in the direction of the Fannie Mae relative sentiment shift, most likely due to contextual biases such as the fact that the distribution will be biased towards finance-related articles and articles published on exactly the same dates. However, since the Fannie Mae sentiment shift gets a value of 0.039, which is greater than any sample statistic, it is highly significant despite these biases. The p-value in each case, as given by Definition 4 is $1e-05$.

It was hypothesised above that narratives about phantastic objects should increasingly exhibit a gradual elimination of doubt and then, eventually, a massive reintroduction of doubt. First an excited phantastic object narrative is created and propagates and then it collapses. The statistically significant shift in relative sentiment surrounding Fannie Mae suggests that it did become a phantastic object in that way and eventually suffer that fate. Further statistical tests are now used to establish what is suggested by visual inspection of Figure 6.1.2, namely that relative sentiment around the phantastic object (Fannie Mae) went through a period of disconnection with an indicator of 'reality' (the Case-Shiller index viewed as the relevant fundamental data) and so exhibited a divided state. This is done using the sample Pearson correlation coefficient.

Figure 6.1.4 shows rolling correlations between the Fannie Mae relative sentiment and the Case-Shiller index with a window size corresponding to 12 periods, i.e., one year. In other words, each bar on the graph represents the sample Pearson r obtained when computed on the indicated 12 months. A critical observation must be made before computing the correlations: as is reflected by the publication date of the index in Figure 6.1.2, the house price index is reported with a 2-month lag, which means that any possible correlation with the relative sentiment would occur with the same lag (since the release date is when news sources can first react to the reported figures). Therefore, when computing the correlations, the Case-Schiller index has been shifted forward in time by 2 months (just as was done when plotting Figure 6.1.2).

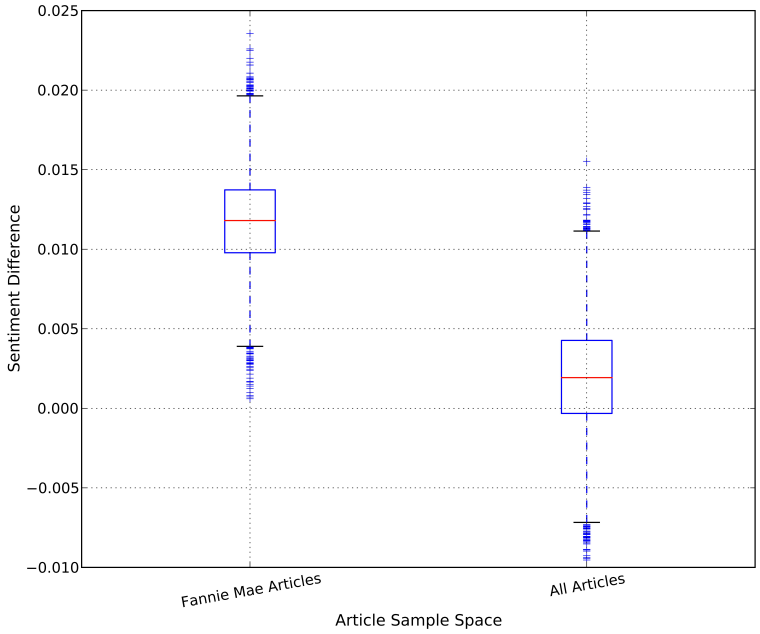

Figure 6.1.3: Box plots of sentiment shift distributions consistent with Fannie Mae RSS; sample space given by Fannie Mae articles vs. all articles

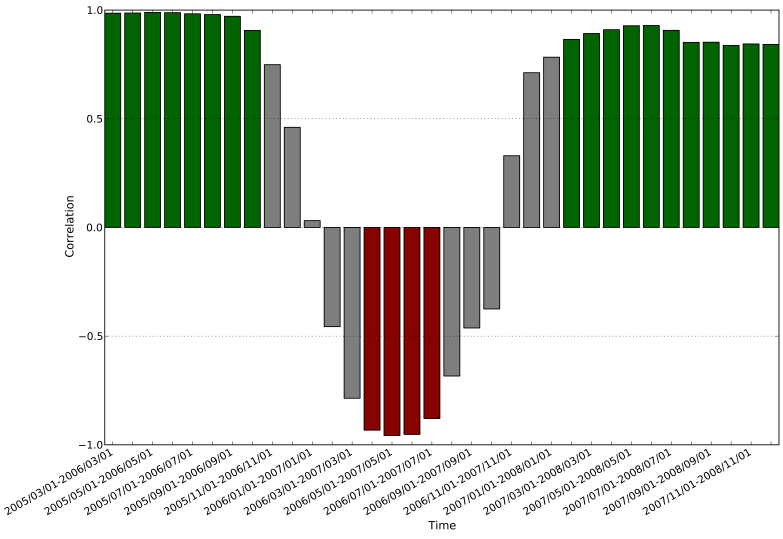

Figure 6.1.4: Correlations between Fannie Mae RSS and Case-Shiller house price index with a 12-month rolling window

The graph portrays periods of those qualitatively significant evaluations of the correlation (red and green), as well as its overall progression through time. The graph clearly shows that relative sentiment was from early 2005 correlated with the house price index, only to become negatively correlated during the critical period (when, apparently, fundamental data were ignored by investors) and finally becoming correlated once again, post crisis. The graph indicates how relative sentiment became disconnected - and, in fact, illogically connected - to fundamentals during the critical period). This is indicative that there was a divided state associated with thinking and writing about Fannie Mae. Fannie Mae likely became a phantastic object apprehended through a divided state during this time period.

This supports the hypothesis that the story surrounding Fannie Mae became a phantastic object experienced in a divided state and that the RSS methodology focused on articles about Fannie Mae can be used to measure the characteristics of such an object.

6.2 Phantastic Object 2: Enron

The analysis of news articles related to Fannie Mae made it possible to detect and characterise the relative sentiment patterns leading up to the critical period as well as how those patterns played out afterwards. However, it was not possible to detect potential differences in the propagation of narratives and shifts spreading across different social networks. Although the Enron database is not historically complete in the same sense as the news archive, it might be possible to use the relative sentiment shift methodology and the email database to explore something not possible in the Reuters data on Fannie Mae - namely the way conviction narratives percolate through networks in an organisation.

To date, the majority of research conducted on this dataset has been in natural language processing, in which the data have been used to train email classifiers that attempt to label emails by some of their attributes (e.g., the email folder, sender or recipient) or by the email content [56], [57]. A further study, linked various network properties of the associated graph with the available historical accounts [32]. The study presented here is distinctly different and can also be found in [114].

The Enron Corporation had a dramatic history [51], [71]. One of its most important business projects was the plan to move from energy wholesale into the energy retail business selling directly to consumers rather than to the regulated utility companies. Had it been allowed Enron claimed it could provide consumers with lower prices and itself make very high profits, by removing the middleman. The plan, which was particularly aimed at the California Energy market, depended on successfully

lobbying for regulatory change for which, in 1999 alone, Enron's government affairs budget exceeded $37 million [71]. Anticipating success, Enron created long-term contracts with consumers and participated in pilot programs with some US states. By the end of 1999, Enron had 'sold' long-term energy contracts to an estimated 'value', (according to Enron) of $8.5 billion [71]. It seems Enron did not anticipate successful counter-lobbying from the utility companies it planned to bypass and had not paid much attention to how its contracts and payments would eventually pay off. The idea was to get "big" customers like IBM and Chase Manhattan Bank. All this (and the eventual collapse) suggested that a narrative had been developed around the benefits to Enron of deregulation of the California retail energy market so that it might have been a phantastic object narrative developed and sustained by a divided state.

A second important and apparently very similar Enron project was chosen to test whether the same patterns would be found as with the energy topic. The project concerned Enron's ambition to enter into the broadband delivery business, a business in which Enron had no previous experience. Jeff Skilling has been recorded as saying,

> "I've always believed there's no such thing as a free lunch, but this looks like a free lunch. I've never seen economics like these numbers."

[71] - a statement reminiscent of many made in the dot-com boom which can be considered an exemplar of thinking in a divided state in which beliefs about reality are over-ridden and contradicted by excitement about a phantastic object narrative [113]. Emails mentioning California Energy and Broadband, therefore, will be examined to test for the characteristic patterns of excitement and doubt.

6.2.1 Hypothesis

From conviction narrative theory it would be expected that both these topics ('retail energy', in particular the California energy market, and 'broadband', henceforth referred to as 'California' and 'Broadband' respectively) would show the characteristics of a phantastic object; divided state and groupfeel. It is conjectured that the sentiment surrounding discussions of the above topics should exhibit significant shifts, in at least some of the important parts of the social network implicit in the emails.

It is hypothesised that some of the key social groups at Enron exhibited a significant shift in sentiment, specifically focused on these phantastic objects, at the height of Enron's misconduct surrounding them.

As with the case study about Fannie Mae in Reuters News, this examination is entirely retrospective in the sense that most information about the history and the

narratives selected for study was available when the topics were selected. However, both studies are important steps towards being able to capture phantastic object narratives as they develop without knowledge of the underlying entities involved. Chapter 7 will explore one methodology that could be used to eventually achieve this.

6.2.2 Methodology

The above hypothesis is tested similarly to how the significance of the trend in the relative sentiment series of Fannie Mae was tested. Given a number of interest (i.e., a particular value for a particular social group during a particular time-period) test if it is a significant value for that period, given the number of emails, and the measure that generated it. An empirical distribution of the potential values will be generated via random sampling (Definition 3) and used to test the hypothesis (Definition 4).

The steps of the analysis may be summarised as follows:

1. Construct a graph with senders as vertices and emails as edges. Use the entire email database to construct this graph.
2. Cluster the network into groups using this graph.
3. Filter all emails, sent by these groups, on the candidate phantastic object(s).
4. Apply the RSS methodology on the selected emails, making sure to scale by the total relative sentiment of the data.[4]
5. Finally, test for significance of the results by applying Definitions 3 and 4.

The deterministic graph-clustering algorithm by Newman [76] is applied on the constructed email network in order to extract social groups based on a 'larger than expected' (as measured by a power law model estimation of the graph) intra-connectivity. There are several reasons for using this particular graph algorithm; its determinism (for replicability), its reported accuracy despite the efficiency of the approximation [76] and the many available programming implementations [3]. The details of the algorithm can be found in Chapter 2 and of course also in the original paper by Newman [76].

[4]This procedure differs somewhat from the generic equation 4.1 and from the approach taken in the previous section. The reason for this is to better distinguish the relative sentiment of the target topics from the normal emotional tone of the email database. This was easy to do with Fannie Mae because the regular structure of a news article made it possible to split it into sentences and focus only on those that matched the pattern, whereas email data often lack the usual grammatical structure of written texts.

The regular expression patterns employed here to filter for the candidate phantastic objects are created from two sets of keywords, relevant to the two business strategies, each set combined into one pattern by the use of logical disjunction ('or' operators).

- In the case of California Energy the keywords used were: 'california', 'retail', 'regulator', 'regulators', 'regulatory', 'regulation', 'deregulation', 'deregulate', 'deregulates', 'deregulated', 'regulated', 'regulate', 'regulates' and 'caiso'. The only word in this list in need of clarification is 'caiso'. This word is a frequently used abbreviation of California Independent System Operator (ISO) - an independent, non-profit organization that oversees the operations of the California power system.
- In the case of the Broadband business strategy, the keywords used were: 'broadband', 'firstpoint', 'bandwidth', 'communications', 'hirko', 'rice', 'network' and 'optics'. In this list there are three words in need of clarification, 'firstpoint' refers to FirstPoint Communications - a startup inside Portland General acquired by Enron to become Enron Broadband. 'hirko' and 'rice' refer to the co-CEOs of Enron Broadband - Joe Hirko and Kenneth Rice.

The relative sentiment of the target topics normalised by the total relative sentiment is formally given by,

$$\frac{Exc(X)}{Exc(T[P])} - \frac{Anx(X)}{Anx(T[P])} \tag{6.2}$$

where T is the total collection of emails and P is a given time period.

6.2.3 Results

Firstly, the relative sentiment series (normalised by the total relative sentiment as given by equation 6.2) of all emails that match the two topics, 'California' and 'Broadband', are computed and plotted with exponential smoothing parameter 0.2 in Figures 6.2.1 and 6.2.2 respectively. For both topics an increase in anxiety relative to that found within all emails can be seen.

Secondly, cluster the full network of emails using the clustering algorithm of Newman [76] as implemented in the IGraph library [3]. The algorithm is forced to stop after it found a maximum of 10 social groups to avoid segregating the data too much. These automatically discovered clusters aligned well with natural groups of high-profile individuals (from the known Enron history), which can be seen in Table 6.2.1.

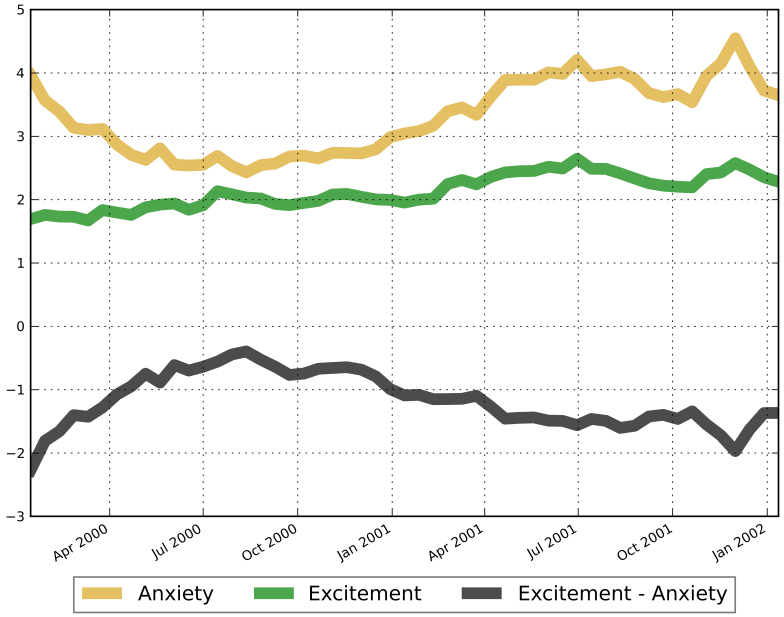

Figure 6.2.1: Enron California RSS, Excitement and Anxiety series

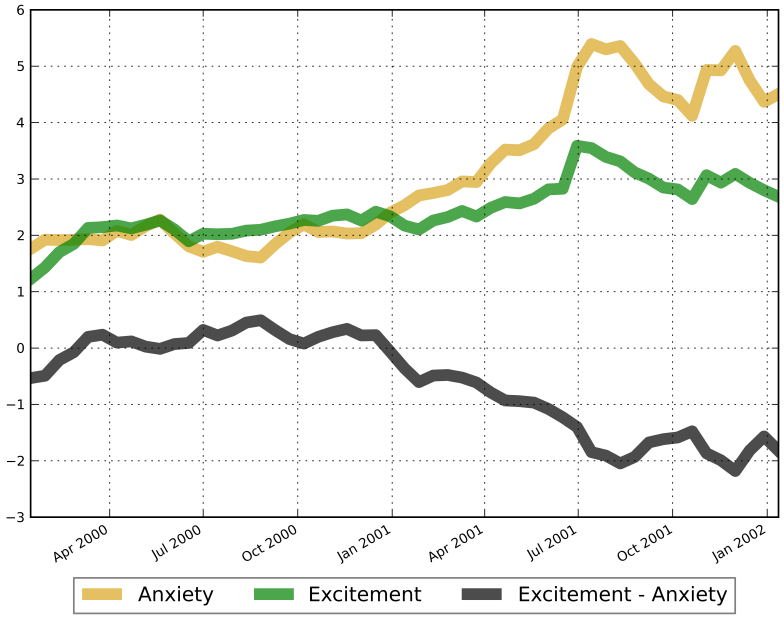

Figure 6.2.2: Enron Broadband RSS, Excitement and Anxiety series

Table 6.2.1: Enron clusters of all 75 558 employees present in the email database

Label	Size	Key Enron staff
0	44130	Rick Causey
1	22195	John Arnold, Rick Buy, Margaret Ceconi, Jim Derrick, Sherron Watkins, Anne Yaeger
2	1402	David Cox
3	1428	Andrew Fastow, Rebecca Mark, Ray Bowen, Rebecca Carter, Wanda Curry, Greg Whalley, Tom White
4	3432	Jeff Skilling, Cliff Baxter
5	1786	Tim Belden, Dave Delainey
6	233	N/A
7	84	N/A
8	755	Ken Lay
9	113	N/A

Note here that even though the clustering algorithm used is entirely dependent on the network structure alone, some similarity can still be found between how the key individuals have been grouped and the underlying Enron story. Focusing on cluster 4, it is known that Skilling's perhaps closest co-worker actually was Baxter, and hence it is perhaps not surprising that these important operational policy innovators are grouped together by the algorithm. Within cluster 3 can be found Chief Financial Officer Andrew Fastow, Enron finance executive and briefly treasurer Ray Bowen and accountant Wanda Curry, among others. For the rest of the study clusters 6, 7 and 9 have been excluded due to the sparsity of data for these clusters.

The objective is to quantify the progression of relative sentiment within each of the 7 social groups of interest, to identify any significant differences between them. In this study the entire available time range has simply been divided into two intuitive halves representing two distinct periods in the developments of the Enron story to quantify the progression of relative sentiment within each social group.

- Let P1 be the time-period from January 1, 2000 until January 1, 2001.
  - This is considered to be the period leading up to the crisis; a period in which anxious information is expected to have been subconsciously rejected.
- Let P2 be the time-period from January 2, 2001 until January 1, 2002.
  - This is considered to be the period when the crisis started to unfold. In

particular, Jeff Skilling took over as CEO on February 12, 2001 and the Enron stock price started its steep decline.

Having defined the time-periods during which a shift is expected to have taken place, a sentiment shift function is defined which makes it possible to formally test for significance of the hypothesis; if a significant shift can be observed in a particular social group compared to the rest of the network.

Definitions 1 and 2 are refined to specify the collection of emails sent from a particular social group.

Definition 5. *For two clusters, M_1 and M_2, define the set of* inter-cluster *emails from cluster M_1 to cluster M_2 conditioned on time-period P by*

$$\begin{aligned} InterCluster[M_1, M_2, P] = \{e[body] | (e \in D) \wedge (e[sender] \in M_1) \\ \wedge (e[receivers] \cap M_2 \neq \emptyset) \wedge (e[date] \in P)\} \end{aligned} \tag{6.3}$$

Notice that $InterCluster[M_1, M_2, P] \neq InterCluster[M_2, M_1, P]$ by convention and that not all receivers are required to be in cluster M_2. Also note how it is possible to remove the condition on the time-period P to construct $InterCluster[M_1, M_2]$ with obvious meaning. It is also possible to further condition the emails by filtering on tokens within the email bodies.

With this definition, define the relative sentiment metric to apply to each social group, M, to quantify its sentiment progression,

$$\begin{aligned} \{\frac{Exc(InterCluster[M, ALL, P|Topic])}{Exc(T[P])} - \frac{Anx(InterCluster[M, ALL, P|Topic])}{Anx(T[P])}) \\ |P \in \{P_1, P_2\}\} \end{aligned} \tag{6.4}$$

where T is the entire text collection as in Definition 1, $InterCluster[M, ALL, P|Topic])$ implies a further condition on the inter-cluster emails by considering only those containing at least one token characterising the topic (Topic = CA, for California or BB, for Broadband). With these definitions in mind it is possible to attempt to quantify a sentiment shift conditioned on emails going from a particular cluster and emails regarding the given topics.

Definition 6. *Let $f(X, Y)$ be the following function, which will be called* sentiment

shift function, *of collections of texts* $X \subseteq T[P_1|Topic]$ *and* $Y \subseteq T[P_2|Topic]$.

$$f(X,Y) = |(\frac{Exc(X)}{Exc(T[P_1])} - \frac{Anx(X)}{Anx(T[P_1])}) \\ -(\frac{Exc(Y)}{Exc(T[P_2])} - \frac{Anx(Y)}{Anx(T[P_2])})| \tag{6.5}$$

In other words, $f(X,Y)$ *computes the absolute value of scaled sentiment difference of emails* X *from period* P_1 *and emails* Y *from period* P_2 *which are filtered on the California or Broadband PO words. Where scaled sentiment, of* $X \subseteq T[P_1|Topic]$*, is defined as* $\frac{Exc(X)}{Exc(T[P_1])} - \frac{Anx(X)}{Anx(T[P_1])}$*. The score is normalised by the sentiment of all emails falling within a particular time-period to remove any sentiment bias present in the database.*

The relative sentiment of inter-cluster emails going from each social group to all the rest in time-periods P_1 and P_2 is plotted in the left hand side column of Figure 6.2.3. Note the dramatic difference between cluster 4 and the other clusters. To test for significance of the relative sentiment shift of cluster 4 relative to the other clusters apply the sentiment shift function defined in equation 6.5 to random samples of emails from the collection of all emails that match a particular topic; "California" or "Broadband", given the same sample size for X and Y as observed for cluster 4. In other words, repeatedly sample from $T[P_1|Topic]$ and from $T[P_2|Topic]$ the same number of emails about the given topic as sent from cluster 4 in periods P_1 and P_2 to generate the distributions for the two target topics.

The two distributions are plotted on the right hand side column of Figure 6.2.3. Notice that the relative sentiment shift of emails classified as "California" sent from Skilling's group, which is assigned a "scaled relative sentiment shift word count difference" of approximately 2.45 words per email, is a very significant result; exceeding all sample values. Likewise, the equivalent score assigned to cluster 4 for the topic of "Broadband" is 4.24 and is also a highly significant value.

Which social group accounted for most of the sentiment shift regarding the two topics has now been identified. It has also been shown that this social group experienced a significant shift relative to the other social groups (holding the topics fixed, i.e. conditioning the sample space on the topics). But how is it possible to tell that the significance of the shift of cluster 4 is really attached to the two topics, and is not actually present in all the emails of cluster 4? In other words, what has really been gained by focusing the study on the "California" and "Broadband" topics in terms of the difference observed between the clusters?

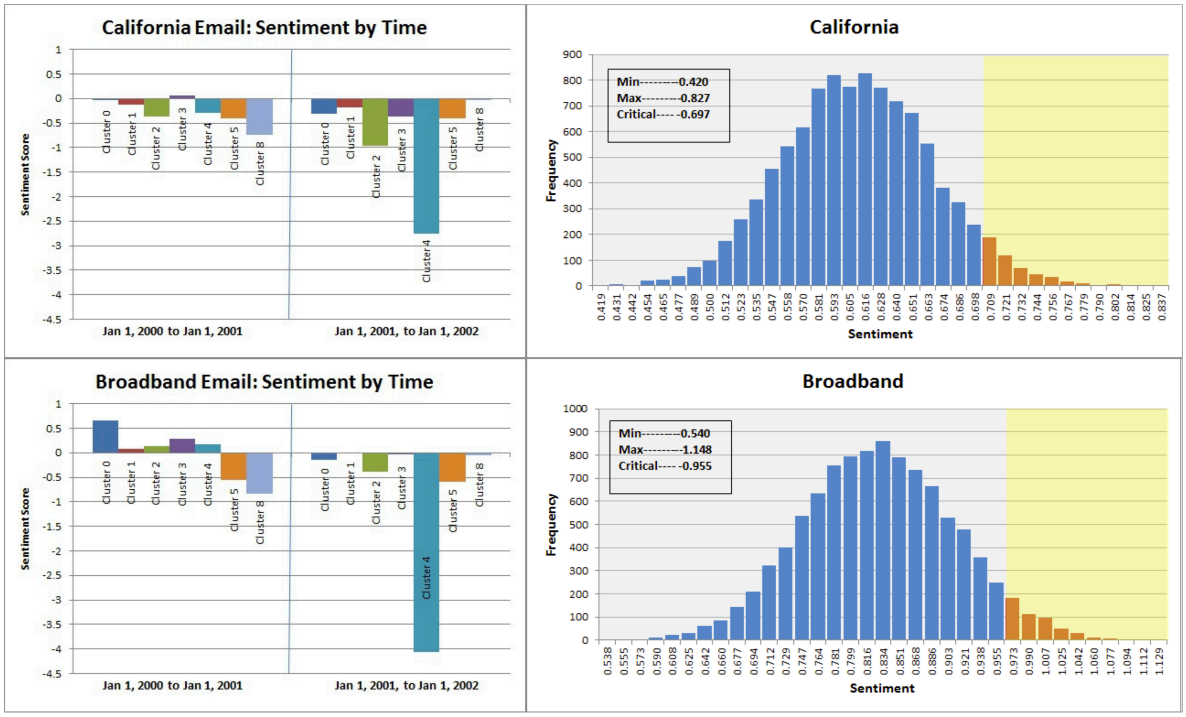

Figure 6.2.3: Enron California and Broadband RSS changes within key social groups

This can easily be tested by looking at the sentiment per cluster found in all non-PO emails; emails which are not classified as either "California" or "Broadband" emails. If the difference between cluster 4 and the others persists then no significance has been gained by focusing the analysis on the PO topics. If however the difference disappears, the significance of the methodology is clear. Figure 6.2.4 shows the corresponding 'sentiment by time' bar chart, displaying the value of the approach by clearly showing the latter.[5]

Returning to the sentiment series of the topics across all emails, shown in Figure 6.2.1 and 6.2.2, the significance testing methodology can be applied to conclude that both topics alone experienced significant shifts in sentiment relative to the baseline of all emails. Thus there are two levels of significance, one given the topics ("California" and "Broadband") relative to all emails and one given the social group of Jeff Skilling relative to the other social groups but conditioned on the topics.

Therefore, clustering the network into social groups made it possible to achieve two things. First of all, it gave a finer level of granularity in which to look for significant shifts in sentiment. Secondly, and most importantly, it made it possible to conclude that most likely there was a difference in emotional states between one group and the rest. Since it is reasonable to expect that all groups had the same information about the "real" state of real markets and overall politics external to Enron, this shift in

[5]Please note the much smaller scale of Figure 6.2.4 when compared to Figure 6.2.3.

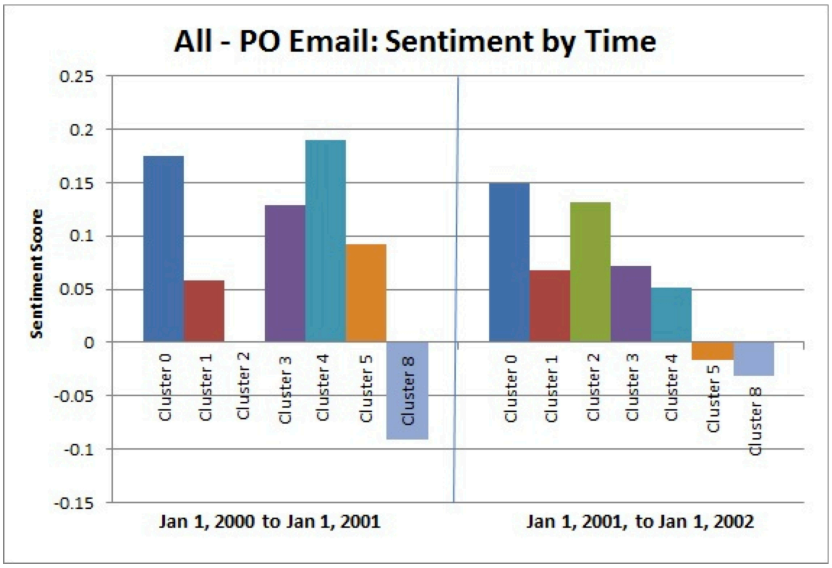

Figure 6.2.4: Enron RSS changes in all emails within key social groups

sentiment is strongly indicative of a divided state. The tie of this significance to (retrospectively) known objects that turned out to be phantastic, lends credence to the main hypothesis: that technical means can be used to operationalise elements of the underlying psychological and sociological theory of conviction narratives.

6.3 Discussion

This chapter has focused on particular narratives of interest, which as indicated by the title of the chapter are referred to as a 'micro' level narrative analysis (to distinguish it from the aggregate 'macro' narrative analysis of the previous chapter). These narratives, the narrative involving Fannie Mae as represented in Reuters news and the two business narratives of Enron as represented in their internal email archive, were chosen retrospectively because they were assumed to show certain characteristics of phantastic object narratives. In particular, it was hypothesised that all narratives would show a diminishing amount of anxiety relative to excitement over a period in which a divided state would have developed. In the case of Fannie Mae, it was possible to show such a significant decline in relative anxiety, by applying a Monte Carlo based approach, over a period in which fundamental information about the 'true' state of the world failed to catch investor attention. This pattern typifies a divided state.

Similar emotional trajectories could be observed for the two Enron narratives, retail energy and broadband. However, in this case the database analysed did not go far back enough to detect a significant decline in anxiety leading up to the tipping point. Neither is there a simple proxy for the real state of the world that clearly contradicted the internal conviction within Enron - except in the form of ex post knowledge of how

both projects eventually turned out to lack any real substance. Nonetheless, what was available through the Enron study, that was not available for the Fannie Mae study, was information about social influence. Using the inherent social network characteristics of the Enron email database, and a methodology for significance testing of sentiment analysis metrics, it was shown how the social group of Jeff Skilling experienced a significant shift in sentiment towards "anxiety" taking place at the time of the height of Enron's stock price, which can also be considered the time when Enron's bad practices were just about to be unravelled. This shift appears to have been unique to this social group and is consistent across the two phantastic object candidates studied; the "California" narrative and the "Broadband" narrative. The fact that the sentiment shift was dependent on the topic (since the sample space was conditioned on relevant emails only, and the sentiment metric was scaled by the sentiment present in the entire data collection) and comparative to other individuals suggests that a divided state about the noted phantastic objects was present in this social group. Furthermore, the consistency across the two phantastic object candidates further strengthens the conclusion that such a divided state was present.

Both case studies support the main ideas of conviction narrative theory and the insights derived could potentially be used to detect such narratives ex ante.

In the following chapter, outlining the third and final experiment of the thesis, a methodology will be developed that connects the two different views on relative sentiment shifts - the macro and micro view - and also sheds some light on how one might go about detecting phantastic object narratives ex ante.

7

Experiment 3: Understanding the Macro from the Micro

This chapter will attempt to relate RSS of particular narratives (the micro view) to the aggregate RSS measure as applied to a collection of narratives (the macro view). The aim, as previously stated, is twofold. First, to better understand, from a qualitative point of view, what happens when the macro RSS series exhibits large shifts. Secondly, to relate the degree of narrative homogeneity, or the formation of a dominant narrative, to the issue of financial stability. It is shown that macro shifts in confidence can typically be linked to a smaller than average number of 'dominant' narratives (i.e., as RSS shifts on a macro scale there is often a dominant story). Various potential measures of narrative homogeneity are further shown to lead measures of financial stability such as the VIX and the S&P500 equity index. The potential risks associated with groupfeel type behaviour have been illustrated by other studies, e.g., when assessing the sentiment of a new vaccine, using Twitter data, it was shown that most communities are dominated by positive or negative sentiment and that this may increase the risk of disease outbreaks [92].

The methodology developed to test these hypotheses can potentially be further developed to detect narratives, or aspects thereof, showing some of the patterns studied in the second experiment, thus be used to detect potential buildup of 'cognitive' risk.

The chapter starts by stating the hypothesis and then moves on to explain the methodology used to test it before the results are presented followed by a conclusion.

7.1 Hypothesis

The first hypothesis is that macro relative sentiment shifts arise exactly when a narrative focus has emerged, emotional homogenisation would therefore be present at this time and the emotions of the group organised around the narrative focus. This may be expressed more technically, and simplified by focusing the hypothesis only on key entities within the narratives, as follows. *The degree of cluster formation in a relatedness network of key entities within the narratives correlates with shifts in relative sentiment (regardless of the direction of that shift, towards excitement or towards anxiety).* In other words, as shifts in the macro relative sentiment series take place there is a tendency for a large proportion of narratives to have become interconnected. As shifts in the US RTRS RSS series tend to be driven by (more or less) anxiety, the exact relationship between the degree of narrative homogenisation and the US RTRS RSS series is hypothesised to be negative, i.e., a negative correlation between the two variables.

The second hypothesis relates to the issue of financial stability. Given the above discussion, the narrative homogeneity measure is expected to correlate with, and might even potentially be used as a forward looking measure of, financial instability.

7.2 Methodology

The two hypotheses will be tested by first constructing a dynamic network representation of the key entities within Reuters news archive, which due to the nature of news, will be considered to be organisations. The methodology will simply define two organisations to be related if they are mentioned in the same sentence at least once during the period of consideration. Therefore, no assumptions are made regarding the nature of that relationship. Thus, a lot of potentially rich information will be disregarded. The focus will be on the US Reuters news database for the same reason as in Chapter 5, that it is likely a more closed economy for which domestic corporations feature more prominently in the news than international ones, and that the effects on financial stability can be more clearly identified.

The General Architecture for Text Engineering (GATE) library [1], a mature java library of text mining tools, is used to extract the entities mentioned within the news stories. In particular, the standard ANNIE processing pipeline is used to perform sentence tokenisation, named entity extraction and co-reference resolution to arrive at

an annotation database of which entities appeared in the text and in what sentence. Thanks to the co-reference module it is possible to discern when two slightly different mentions of an organisation within the same document likely refer to the same entity. However, co-reference resolution only works on the document level, so if an entity is referred to in different ways in different documents (e.g., different abbreviations etc.) all the different forms will be present in the final list. For more details about the specific algorithms used please see Chapter 2.2, Technical Review.[1] This process is performed on all documents in the US Reuters news database.

Once the annotation database has been constructed a dynamic graph representation of entity sentence co-occurrence is created, and implemented in the graph database Neo4J [5]. With this graph database it is possible to store graph metadata, such as the date of a relationship (an edge) between all the entities found in the Reuters database. This allows for a simple construction of time series of graphs with varying frequencies from standard database queries.

A lower bound on the number of times an organisation (or more specifically, a variation on its name) must be mentioned during the course of a single period, to be added to the list of nodes for that period, is fixed in advance. Such a threshold helps to reduce noise and creates a more manageable graph size. In this particular example an arbitrary lower bound of 5 mentions is set. Again, the point of this exercise is not to optimise on this particular threshold, so this number will remain fixed throughout. All organisations that are not connected to at least one other organisation (i.e., each node included must have at least one neighbouring node) are excluded, as they will make no difference in the final analysis.

A collection of 2 sample months in 2008 (August and September) is considered to illustrate the procedure. Table 7.2.1 shows a list of all organisations the system is able to extract and their corresponding frequencies during these two particular periods. Notice that the system appears to extract sensible organisations.

Notice that Reuters itself is included as one of the organisations extracted. Reuters started displaying their own name in articles around the end of 1997. Henceforth Reuters is excluded from the analysis for the simple reason that including it would likely bias the measures in unintended ways. After all, the objective is to measure links between organisations represented through narratives, and not links due to editorial style.

Figure 7.2.1 displays a subset of the network for a single period. The Figure is

[1]The specific method of named entity extraction used is not the focus of the experiment and as long as the resulting list of organisations found, and the corresponding network that is derived, appear sensible the methodology will be considered successful for the purpose of testing the hypotheses.

Table 7.2.1: The 30 most frequent organisations extracted from US RTRS (frequency in parenthesis)

August 2008	September 2008
reuters (1132)	reuters (1838)
fannie mae (683)	freddie mac (1040)
freddie mac (666)	fannie mae (1030)
nasdaq (157)	morgan stanley (437)
us federal reserve (94)	bank of america (291)
citigroup (84)	senate (277)
us agriculture department (76)	wachovia (249)
morgan stanley (73)	republican (232)
boeing (72)	lehman brothers (218)
pentagon (71)	congress (208)
lehman brothers (59)	treasury (206)
treasury (57)	merrill lynch (206)
ba (48)	goldman sachs (204)
washington post (45)	citigroup (192)
goldman sachs (43)	white house (172)
us treasury (42)	american international group (167)
white house (41)	the fed (149)
general electric (41)	us federal reserve (142)
nyse (41)	house (138)
merrill lynch (38)	nasdaq (135)
bank of america (37)	goldman sachs group (108)
sec (37)	house of representatives (107)
congress (35)	us agriculture department (101)
us treasury department (35)	federal reserve (92)
wal-mart stores (34)	washington mutual (89)
qualcomm (32)	sec (86)
ups (31)	us treasury department (82)
new york stock exchange (31)	boeing (80)
american international group (30)	us house of representatives (73)
usda (29)	us treasury (67)

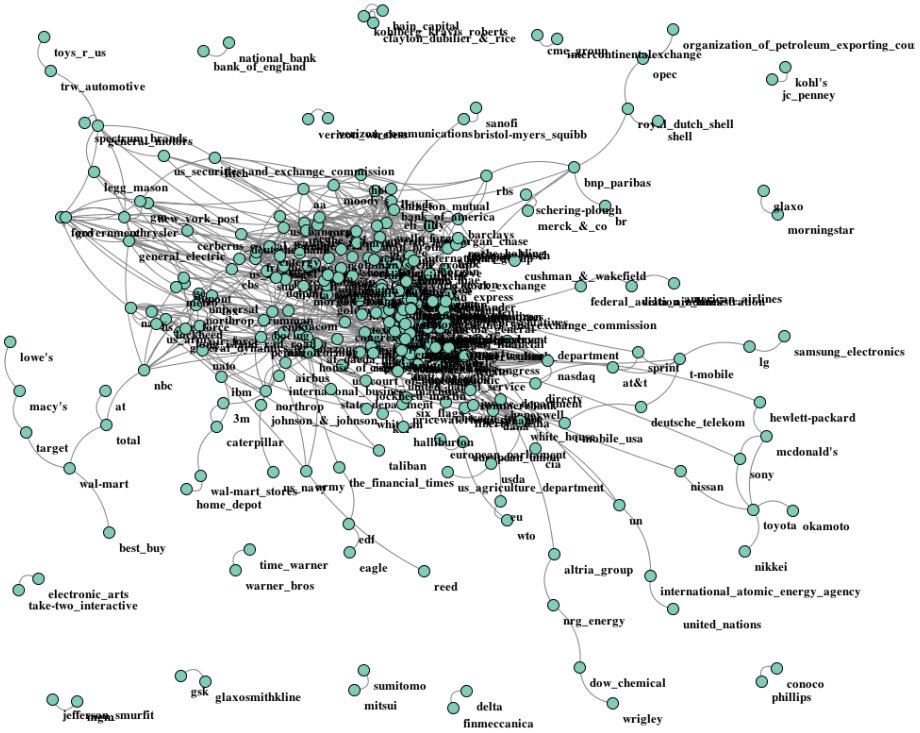

Figure 7.2.1: Organisation co-occurrence network; September 2008

generated using the Jung graph library available in Java [4].

Several network measures will be applied on each individual period of the dynamic network, to create different time series representations of the degree of cluster formation of US organisations. In particular, the graph density, the clustering coefficients and the relative sizes of graph clusters will be considered (in particular how these measures vary over time). The graph density measure will tell how connected the organisational link graph is, by weighting all links equally. The clustering coefficient measure will tell how strongly the graph clusters into different groups, by measuring the degree of node transitivity. It is expected that these two measures are high when there is a shift in macro relative sentiment and during times of financial instability.

It is also possible to abstract away from the links between nodes and simply cluster the graph into groups of nodes and measure when a smaller subset of the graph clusters are relatively much bigger than the rest (i.e., when dominant clusters emerge). It would be expected that this measure also increases during times of relative sentiment shifts. All three of these measures could potentially pick up on the formation and dissipation of a narrative focus on a macro scale.

These measures will be tested for correlation with the macro RSS measure derived for the US. They will also be compared to variables related to financial stability such

as the VIX and changes in the S&P500 index.

7.2.1 Alternative Methodologies

Several alternative means to measure narrative homogeneity can be imagined. For example, semantic models such as Latent Semantic Analysis (LSA) could be used to get a semantic representation of the articles and relate these representations. A similar network could be constructed using this technique or they could be used together with clustering techniques (that automatically derive the number of clusters, for example X-means [86]) to group them into semantic themes. Methods along these lines have been developed to assess the emergence of narrative consensus [77]. However, since the purpose here (partially) is to gain more qualitative insights into what is driving the macro relative sentiment index, output from this type of approach tends to be slightly harder to interpret than the entity network constructed here.

Another potential branch of useful models are to be found in the family of topic models, such as Latent Dirichlet Allocation (LDA). These type of models are becoming increasingly popular in the social sciences and particularly in the field of economics. For example LDA has been used to quantify discussions by different members of the Federal Open Market Committee [48] to assess different theories of deliberation during committee meetings. Other research areas have also seen useful applications of topic models, e.g., LDA has been used to aid in the understanding of large collections of computer source code [69] and for semantic annotation of satellite images [62] to name a few. Topics derived via applications of LDA models tend to be easier to interpret than what is possible to derive from semantic models. There are methods to infer the number of topics to model. However, it can be argued that this approach is slightly convoluted for the purpose of testing the above hypothesis and that the entity network approach is far more parsimonious.

Furthermore, both these alternative methodologies require substantially more computing time and despite currently being very popular they will not be considered further here.

7.3 Results

7.3.1 Measures of Dominant Narrative Formation (DNF)

To test the two hypotheses, several measures referred to as *dominant narrative formation* (or DNF for short) will be defined.

Graph density A ratio of the number of edges to the number of vertices. Higher graph density would imply that organisations tend to co-occur more often. This

measure does not take into account the structure of the edges, in other words *how* they are connected. After having constructed the graphs for each period, this measure has no further parameters. This will be referred to as *Density*

Clustering coefficient A measure of edge transitivity, i.e., of the probability that two vertices are connected given that they are both connected to some third vertex. This measure does take into account the structure of the edges. A high clustering coefficient would imply that clusters are more likely to form and that the number of edges needed to traverse from one vertex to another is small. After having constructed the graphs for each period, this measure has no further parameters. This will be referred to as *Cl. Coeff.*

Entropy of graph cluster membership distribution A measure of the degree of predictability of cluster membership. A lower entropy would imply that there are more dominant vertex clusters (given a vertex it is easier to predict which cluster it belongs to). This measure is perhaps most similar in nature to a measure of 'narrative dominance'. To construct this measure two parameters need to be set. Girvan and Newman's Edge-Betweenness algorithm described in [43], as implemented in JUNG, will be used to cluster the graph. The algorithm has been used extensively and can be considered useful and robust, for example to investigate biological function in protein interaction networks [35]. The first parameter determines the number of edges to remove from the full graph of each period, K. Since each graph has a different total number of edges, this parameter is defined as a proportion of the total number of edges. The parameter is set to half the total number of edges, $K = \frac{1}{2} \times |E|$. The second parameter determines the number of resulting clusters to consider for the entropy measure. After running the clustering algorithm it is found that each period contains a large number of very small clusters (of pairs or even single vertices). Therefore only the N largest clusters in each period are considered. This parameter is set to 10, $N = 10$. Both parameters will remain fixed during the course of the analysis.[2] This will be referred to as *Cl. Entropy*

Figure 7.3.1 shows the total number of vertices and edges found in each month over the full period, normalised by the total number of words in the given period (in other words, in the same way the emotion word frequencies to derive the US RTRS index were normalised). Both the raw series (i.e., without normalisation by total word count) correlate with the total number of articles (vertex frequencies by 0.45 and edge

[2]The actual values are likely to be unimportant for the purpose of comparing the change in the measure overtime with RSS and other variables.

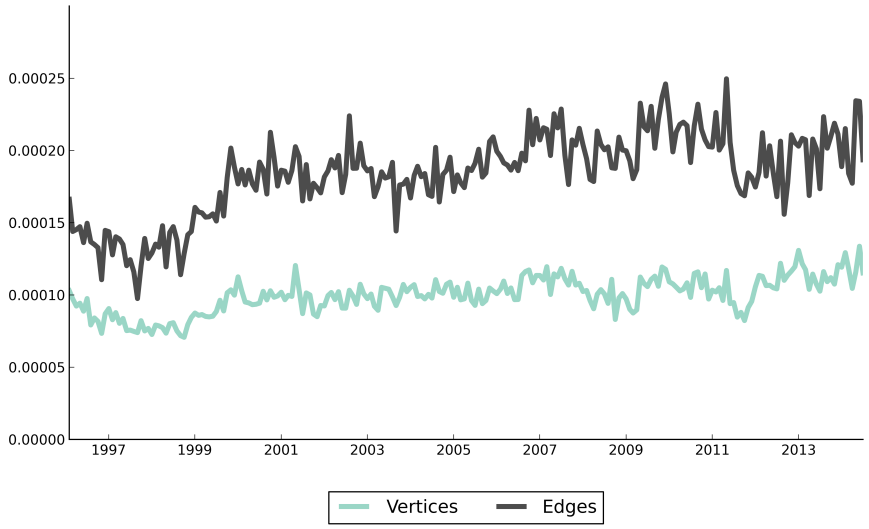

Figure 7.3.1: Scaled (by number of words) number of vertices and edges in dynamic US graph; January 1996 through June 2014

frequencies by 0.28, and the first order differences by 0.72 and 0.74 respectively). In other words, the 'density' of entities mentioned (proportion of mentions relative to all words) in the news as opposed to the raw number is the interesting variable.

For both series, there appear to be an increasing trend. The proportion of different organisations mentioned tends to increase and the proportion of connections between them also tends to increase. Recall that all organisation names not mentioned at least 5 times in a given month are excluded and only nodes with at least one neighbour are included. Neither series appear to show any particularly interesting patterns, they are both relatively 'flat' around the period of the financial crisis. The variability of the normalised frequency of edges is slightly higher than the variability of the normalised number of vertices (the coefficient of variation of the former is 0.156 and of the latter is 0.124). This is in line with expectations, as the number of unique organisations is likely to vary less than the number of links between them.

Figure 7.3.2 shows the three time series created as above, with a monthly frequency, over the full period. In this case, neither of the three series are normalised by the total size of the text in each period. This is because, in each case, the correlations between the first order differences of the normalised series and the total number of words are much greater than the correlations when using the raw series.[3] In other

[3]The correlation between the first differences of normalised vs. raw density and number of words is -0.66 and -0.30 respectively. The correlation between the first differences of normalised vs. raw clustering coefficient and number of words is -0.53 and -0.05 respectively. The correlation between the first differences of normalised vs. raw entropy and number of words is -0.87 and -0.18 respectively.

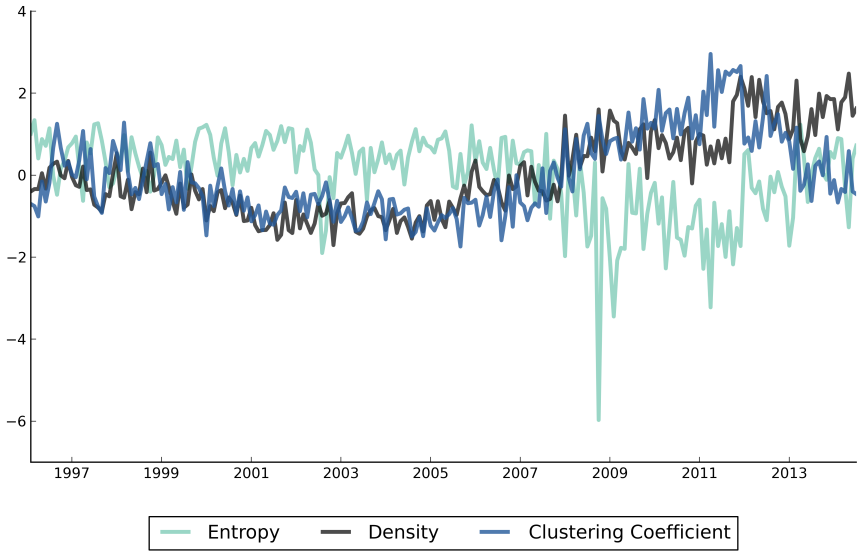

Figure 7.3.2: Dominant narrative formation measures of US graph; January 1996 through June 2014

words, normalising the variables would introduce an undesirable dependency on the size of the text.

To check how all series relate to one another simple pairwise correlation tests are performed and the correlations are reported in Table 7.3.1.

Table 7.3.1: Correlations between DNF measures (p-value in parenthesis)

	Edge Freq.	**Density**	**Cl. Coeff.**	**Cl. Entropy**
Vertex Freq.	0.81 (2e-16***)	0.27 (5e-05***)	-0.02 (0.75)	-0.04 (0.60)
Edge Freq.	1	0.29 (1e-05***)	0.16 (0.02**)	-0.33 (4e-07***)
Density		1	0.70 (2e-16***)	-0.49 (7e-15***)
Cl. Coeff.			1	-0.69 (2e-16***)
Note:		*p<0.1; **p<0.05; ***p<0.01		

Correlations between first order differences of all variables are presented in Table 7.3.2.

Table 7.3.2: Correlations between first order differences of DNF measures (p-value in parenthesis)

	ΔEdge Freq.	**ΔDensity**	**ΔCl. Coeff.**	**ΔCl. Entropy**
ΔVertex Freq.	0.60 (2e-16***)	-0.41 (2e-10***)	-0.24 (0.0003***)	0.30 (7e-06***)
ΔEdge Freq.	1	0.001 (0.99)	0.03 (0.63)	-0.06 (0.34)
ΔDensity		1	0.41 (2e-10***)	-0.32 (2e-6***)
ΔCl. Coeff.			1	-0.49 (1e-14***)
Note:		*p<0.1; **p<0.05; ***p<0.01		

Unsurprisingly, the vertex frequency and edge frequency measures are highly correlated. As the number of organisations that are mentioned relative to the size of the text increases, so does the tendency for them to be mentioned in the same sentence, in other words to be related. It can be noted that, on average, for each vertex there appears to be 2 edges. Indeed, a Kolmogorov-Smirnov test of distributional equality of the first order differences of the normalised number of edges and the normalised number of vertices multiplied by a factor of 2 cannot be rejected (p-value given by 0.90). Although the level of vertex frequency is somewhat positively correlated with density, the first order differences are negatively correlated. Since both the vertex frequency and the edge frequency appear to be trending, Pearson correlation is not well-defined (since the variances of the variables are not necessarily finite), so the correlations between the non-differenced series should probably be disregarded in this case (this does not appear to be much of a problem for the three DNF variables where the act of differencing does not change the sign of the correlation coefficients), or at least treated with care. Recalling the formula for graph density,

$$\frac{2|E|}{|V|(|V|-1)},$$

where $|E|$ is the (unnormalised) number of edges and $|V|$ is the (unnormalised) number of vertices, negative correlations between the unnormalised difference series make sense.[4] As the number of vertices increase, the graph density should tend to

[4]In fact, the correlation between the first order differences of the raw vertex counts and density is -0.64 and the correlation between the first order differences of the raw edge counts and density is -0.24.

decrease since the denominator is quadratic whereas the nominator is only linear and the number of edges is similar in distribution to the number of vertices (scaled by a linear factor of 2, this is true also for the unnormalised series). In other words, if $|E|$ is replaced by $2|V|$, the density function simplifies to $\frac{4|V|}{|V|(|V|-1)}$ which is a decreasing function of $|V|$. The density function also decreases much faster as a function of $|V|$ than as a function of $|E|$ (which reflects the smaller negative correlation with respect to the number of edges). The act of normalising the series changes the correlations somewhat.

Similar correlations between the vertex and edge frequencies and the clustering coefficient and entropy are found (although, for entropy the sign of the coefficient is reversed). In these cases, the interpretations are not as clear and to avoid getting off topic (and potential ex post theorising), this will not be delved into further, other than to note that neither the clustering coefficient nor the entropy measure appear to correlate strongly with the vertex and edge frequency measures. In other words, these measures clearly capture information that differs from the simpler measures.

7.3.2 US RTRS DNF Related to RSS

To begin with, the first hypothesis is tested, whether or not the DNF measures correlate with RSS, as well as the individual emotional components of excitement and anxiety. The results of each Pearson correlation test can be found in Table 7.3.3

Table 7.3.3: Correlations between DNF measures and RSS (p-values in parenthesis)

	EXC	**ANX**	**Density**	**Cl. Coeff.**	**Entropy**
RSS	0.36 (3e-08***)	-0.97 (2e-16***)	-0.52 (2e-16***)	-0.56 (2e-16***)	0.52 (2e-16***)
EXC	1	−0.14 (0.03**)	0.04 (0.51)	0.02 (0.72)	0.19 (0.005***)
ANX		1	0.56 (2e-16***)	0.60 (2e-16***)	-0.51 (5e-16***)

Note: *p<0.1; **p<0.05; ***p<0.01

All three DNF measures correlate significantly with RSS. Anxiety appears to correlate particularly strongly with the three DNF measures, confirming the belief that narrative homogenisation should correlate more with shifts in anxiety because this is the key emotional driver of US RTRS RSS. The correlations between the first

order differences of all variables and corresponding p-values are presented in Table 7.3.4.

Table 7.3.4: Correlations between first order differences of DNF measures and RSS (p-values in parenthesis)

	ΔEXC	**ΔANX**	**ΔDensity**	**ΔCl. Coeff.**	**ΔEntropy**
ΔRSS	0.55 (2e-16***)	-0.96 (2e-16***)	-0.22 (0.001***)	-0.14 (0.04**)	0.23 (0.0006***)
ΔEXC	1	-0.28 (2e-05***)	-0.07 (0.30)	−0.04 (0.52)	0.13 (0.05*)
ΔANX		1	0.23 (0.0007***)	0.14 (0.03**)	-0.22 (0.001***)

Note: *p<0.1; **p<0.05; ***p<0.01

The first order differences of all three DNF measures correlate significantly with shifts (i.e., first order differences) in RSS. The correlations are highest for the density and entropy measures. All three measures correlate significantly with changes in anxiety. Again, the correlations are highest for the density and entropy measures. Only the entropy measure correlates with changes in excitement, although the correlation is very weak and is potentially spurious (even though the p-value is indicating significance at the 5% level, given the number of tests performed, it should be interpreted with a certain degree of scepticism). It can be concluded that when dominant narratives, involving organisations, emerge there is a clear tendency for negative shifts (as indicated by the signs of the correlations) in RSS to occur contemporaneously. Thus, whenever a negative shift in RSS occurs, it is likely that a dominant narrative has emerged. In general, the RSS index appears to be driven by changes in anxiety which is likely the reason why dominant narratives do not appear to emerge when there is a relative increase in excitement; effectively because such an increase is simply due to a *decrease* in anxiety.

This evidence clearly supports the first hypothesis, that macro relative sentiment shifts tend to arise when a narrative focus has emerged.

The second hypothesis, whether or not the DNF measures can be used as leading indicators of financial instability, will now be tested.

7.3.3 US RTRS DNF Related to Financial Stability

The second hypothesis, whether or not the DNF measures are leading indicators of financial instability, will now be tested. The VIX and the first order differences of S&P500 equity index will be considered as the dependent variables. The hypothesis is explored by testing for a leading relationship with these indices.

Table 7.3.5 displays the correlations between first order differences of the DNF measures at time $t-1$ and the VIX and the S&P500 index at time t (in other words, the DNF measures are correlated with the following month's VIX and S&P500 index). Although the correlations are fairly low, they are judged significant by the corresponding p-values and the sign of the coefficients are as expected. For example, decreasing Entropy indicates the emergence of a dominant narrative which is expected to imply an increase in the VIX and a decrease in the S&P500 index.

Table 7.3.5: Correlations between first order differences of DNF measures at time $t-1$ and the VIX and S&P500 index at time t (p-value in parenthesis)

	ΔVIX_t	$\Delta S\&P_t$
$\Delta Density_{t-1}$	0.06	-0.13
	(0.37)	(0.06*)
$\Delta Cl.Coeff_{t-1}$	0.11	-0.17
	(0.11)	(0.01***)
$\Delta Entropy_{t-1}$	-0.19	0.17
	(0.005***)	(0.01***)
Note:	*p<0.1; **p<0.05; ***p<0.01	

Granger-Causality

Potential predicability of a variable given the DNF measures can be tested for using the standard Granger-causality methodology. All three DNF measures are considered as potential explanatory variables. The significance level might usefully be adjusted by a factor of $\frac{1}{3}$.

Table 7.3.6 displays the relevant test statistics for the order of integration of the three DNF variables. The tests establish that all three variables have order of integration equal to one.

Table 7.3.7 displays the Akaike Information Criteria for the respective VAR models in tests involving the VIX, with the chosen lag order printed in bold. In the case of

Table 7.3.6: Augmented Dickey-Fuller and Kwiatkowski-Phillips-Schmidt-Shin tests for stationarity of DNF series, Density, Cl. Coeff and Entropy

Variable	ADF	p-value	KPSS	p-value
Density Level	-2.31(6)	0.44	3.69(3)***	<0.01
Density Diff	-7.43(6)***	<0.01	0.05(3)	>0.1
Cl. Coeff Level	-1.42(6)	0.82	2.13(3)***	<0.01
Cl. Coeff Diff	-9.02(6)***	<0.01	0.04(3)	>0.1
Entropy Level	-2.26(6)	0.47	2.32(3)***	<0.01
Entropy Diff	-7.76(6)***	<0.01	0.03(3)	>0.1

Note: *p<0.1; **p<0.05; ***p<0.01

Density, 4 lags lack serial correlation of residuals and has an AIC score very close to the minimum score when using 6 lags, so this model is selected. In the case of Entropy, 8 lags are needed before there is no evidence of serial correlation (at the 10% level) in the residuals, so this model is selected.

Table 7.3.8 displays the Akaike Information Criteria for the respective VAR models in tests involving the S&P500 index, with the chosen lag order printed in bold. In each case, the number of lags corresponding to the smallest AIC score is used.

Table 7.3.9 displays the test statistics for serial correlation of residuals given the respective VAR models in tests involving the VIX and the S&P500 index.

Table 7.3.10 displays the Wald statistics and corresponding p-values of the Granger non-causality tests of all relevant pairs. Only the Entropy series shows evidence of Granger causality of the VIX, with some evidence in the converse direction. All DNF measures are Granger-causing the S&P500 index (the levels of the index), and only the model with the Entropy index shows evidence of Granger causality in the converse direction.

Given the results of the Granger causality tests there is some support for the second hypothesis, that the DNF measures are forward looking indicators of a weakening equity market and increased volatility.

Table 7.3.7: Akaike Information Criteria from 1 through 15 lags; VAR models with VIX and DNF measures (January 1996 through June 2014)

Lags	VIX & Density	VIX & Cl. Coeff.	VIX & Entropy
1	-10.577	1.855	2.460
2	-10.784	1.513	2.341
3	-10.914	1.450	2.299
4	**-10.918**	1.486	2.280
5	-10.897	1.514	2.313
6	-10.921	**1.442**	2.226
7	-10.913	1.454	2.233
8	-10.884	1.454	**2.231**
9	-10.875	1.481	2.233
10	-10.849	1.505	2.238
11	-10.830	1.540	2.241
12	-10.836	1.556	2.261
13	-10.820	1.545	2.297
14	-10.802	1.570	2.313
15	-10.791	1.601	2.315

Table 7.3.8: Akaike Information Criteria from 1 through 15 lags; VAR models with S&P500 and DNF measures (January 1996 through June 2014)

Lags	S&P & Density	S&P & Cl. Coeff	S&P & Entropy
1	-5.744	6.726	7.357
2	-5.967	6.358	7.230
3	-6.073	6.320	7.201
4	-6.084	6.348	7.163
5	-6.061	6.333	7.168
6	**-6.111**	6.237	**7.056**
7	-6.104	6.266	7.073
8	-6.073	**6.237**	7.078
9	-6.067	6.275	7.080
10	-6.054	6.291	7.104
11	-6.025	6.312	7.123
12	-6.028	6.314	7.150
13	-6.031	6.303	7.178
14	-6.000	6.324	7.181
15	-5.981	6.350	7.202

Table 7.3.9: Portmanteau & Breusch-Godfrey test statistics for serial correlation of residuals; VAR models involving the VIX, S&P500 and DNF measures (January 1996 through June 2014)

VAR model	Portmanteau	d.f.	Breusch-Godfrey	d.f.
VIX/Density	59.52	48	26.67	20
VIX/Cl. Coeff	38.57	40	24.79	20
VIX/Entropy	36.37	32	22.88	20
S&P/Density	47.46	40	21.04	20
S&P/Cl. Coeff	33.92	32	18.94	20
S&P/Entropy	40.59	40	23.37	20

Table 7.3.10: Wald test statistics of Granger non-causality tests between VIX and S&P500 index and DNF measures (January 1996 through June 2014)

Direction	Chi-Sq	d.f.	p-value
Density → VIX	5.6	4	0.24
VIX → Density	5.0	4	0.29
Cl. Coeff → VIX	5.0	6	0.54
VIX → Cl. Coeff	6.0	6	0.42
Entropy → VIX	27.7	8	0.0005***
VIX → Entropy	14.4	8	0.072*
Density → S&P	16.8	6	0.01***
S&P → Density	9.5	6	0.15
Cl. Coeff → S&P	20.9	8	0.0075***
S&P → Cl. Coeff	12.6	8	0.13
Entropy → S&P	19.5	6	0.0034***
S&P → Entropy	14.4	6	0.026**
Note:			*p<0.1; **p<0.05; ***p<0.01

7.3.4 Understanding Shifts in Macro RSS

The network model can be used to qualitatively understand large shifts in the macro RSS series. This section will illustrate how visualisations of network clusters, and the associate RSS scores attached to each organisation in a cluster, can be used to determine the main reasons for a shift in macro RSS. The procedure will be illustrated on data from September 2008, a significant month in which a dominant anxious cluster about the ongoing problems within the financial sector would be expected to have emerged.[5]

Using the edge betweenness clustering algorithm that was used to derive the Entropy series, the full network is clustered by removing half the edges in the network. The 10 largest clusters have sizes 5, 5, 6, 6, 7, 7, 7, 9, 12 and 29 respectively. The organisations of the three largest clusters are,

1. deutsche bank, freddie mac, american international group, treasury, goldman sachs group, goldman sachs, merrill lynch, jpmorgan chase, cnbc, the fed,

[5]Although the selection of this particular month is inherently ex-post because it is now known to have been a significant period, because the entropy measure drops dramatically in this particular month it is not unreasonable to assume that this would have been sufficient reason to explore the most dominant clusters in detail in early October 2008.

congress, nyse, new york times, jp morgan, treasury department, wachovia, barclays, sec, washington mutual, morgan stanley, us treasury, lehman brothers, wall street journal, wells fargo, federal reserve, bank of america, citigroup, fdic, fannie mae

2. boeing, northrop grumman, raytheon, ba, us air force, air force, us army, general dynamics, lockheed, airbus, northrop, navy

3. cox, usa today, cnn, house of representatives, republican, abc, us congress, us senate, white house

The three largest clusters, in the list above, are sorted from largest to smallest. The largest cluster is clearly a cluster of financial institutions, as expected. The second largest cluster appears to be a mix of airlines and the US military. The third largest cluster is a public sector and news cluster.

The relations between the organisations are visualised for the largest cluster and the nodes and edges are scaled based on frequency of occurrence (in total for the organisations to make up the weight of the vertices, and the number of sentences two related organisations have co-occurred in to make up the width of the edges).

Given the rather skewed distribution of frequency of mentions of organisations the size of a vertex in the graph is scaled logarithmically making use of the largest and smallest frequencies across all organisations over a specified benchmark period. Similarly, the width of an edge between two organisations is scaled making use of the largest and smallest width present across all organisations. In other words, the scaled vertex size is given by,

$$10 \times log_{10}(10000 \times \frac{s - s_{min}}{s_{max} - s_{min}} + 1) + 4 \tag{7.1}$$

where s is the size of the given vertex, s_{min} and s_{max} are the smallest and largest sizes of any vertex in the benchmark period. The scaled edge width is given by,

$$10 \times \frac{w - w_{min}}{w_{max} - w_{min}} + 1 \tag{7.2}$$

where w is the width of the given edge, w_{min} and w_{max} are the smallest and largest widths to be found between any two organisations in the benchmark period. The scaled relative sentiment score is given by,

$$\frac{RSS - RSS_{min}}{RSS_{max} - RSS_{min}} \tag{7.3}$$

where RSS is the relative sentiment score of the given vertex, RSS_{min} and RSS_{max} are the smallest and largest relative sentiment scores of any vertex in the benchmark

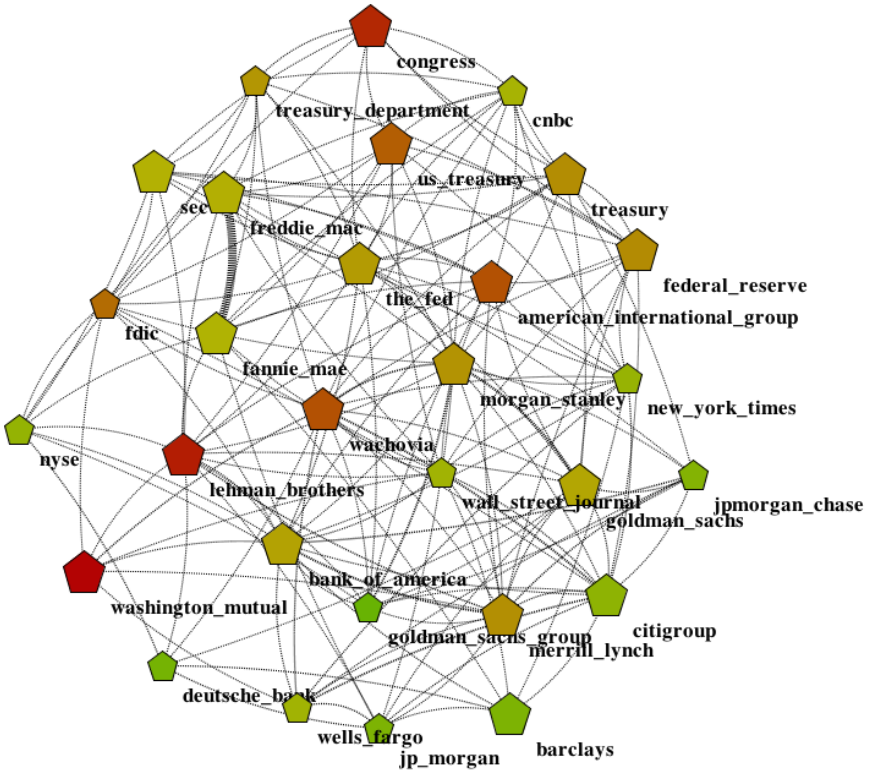

Figure 7.3.3: US organisation network restricted to the largest cluster; September 2008

period after excluding the smallest 2.5% of sentiment scores and the largest 2.5% of sentiment scores (these are excluded simply because a certain number of very large values and very small values are to be expected if the denominator, the total number of sentences an organisation appears in, is very small). The scaled relative sentiment score is then used to generate a colour for the vertex using the Java colour object,

```
Color.getHSBColor(scaled_sent*0.33f, 1f, 0.7f) //scaled_sent represents
    the scaled relative sentiment score for the given organisation
```

Figure 7.3.3 shows the network of nodes in the largest cluster in September 2008, for which the benchmark period for the scaling functions is selected to be January 2006 through September 2008. The figure is generated using the Jung library and the vertex size, edge width and vertex colour are provided to a *renderingContext* object which sets the properties according to the above transformations. The figure shows that the strongest connection is between Fannie Mae and Freddie Mac. The organisations with the most anxious relative sentiment at the time is Lehman Brothers, Washington Mutual, the US Congress, Watchovia, the US Treasury and American International Group (AIG). Furthermore, the network appears highly

connected.

Although this figure is constructed purely for the purpose of illustration, it does show fairly clearly how the negative shift in September 2008 can largely be accounted for by a shift in relative sentiment towards anxiety regarding the banking sector with the expected organisations (AIG, Lehman Brothers, etc.) as the root source.

Figure 7.3.4 and 7.3.5 display the equivalent networks for the second and third largest clusters respectively. Note that the US Congress once more appears, this time 'us congress' has been identified in the text as distinct from 'congress'. However, it is comforting to find that the relative sentiment for this node is equally anxious.

The networks are much smaller and less connected than the largest cluster, represented by the financial institutions. This type of analysis could potentially be refined to capture clusters, and how they evolve dynamically, that exhibit characteristics of phantastic object narratives, i.e. a steady decline in relative anxiety.

7.4 Discussion

This chapter has explored a new methodology to link the narrative topology of organisations mentioned in Reuters news archive with the overall relative sentiment. It was hypothesised that at times of large macro shifts in RSS there will be a corresponding shift towards a single narrative. A methodology was derived that measured the degree of formation of a dominant narrative in several ways by making use of network theory. The density of the graph, the global clustering coefficient of the graph and a cluster entropy measure were considered. These were show to correlate significantly with changes in the macro RSS measure. The cluster entropy measure showed very clearly the formation of a dominant cluster emerging in September 2008, consisting of the major financial institutions.

It was further hypothesised that the dominant narrative formation (DNF) measures could be used as early warning signals of financial instability. This is tested by the application of Granger-causality tests between the candidate DNF variables and the VIX index and the S&P500 equity index. The Entropy measure was shown to Granger-cause both the VIX and the S&P, albeit with some evidence of Granger-causality in the reverse direction. The Density and Cl. Coeff. indices were shown to Granger-cause the S&P500 index with no evidence of reverse causality.

There is therefore substantial support that both hypotheses stated in the beginning of this chapter are in fact true. Considered together, this implies that society can at times become organised emotionally around a particular narrative, and when this happens it can have serious negative implications for financial stability.

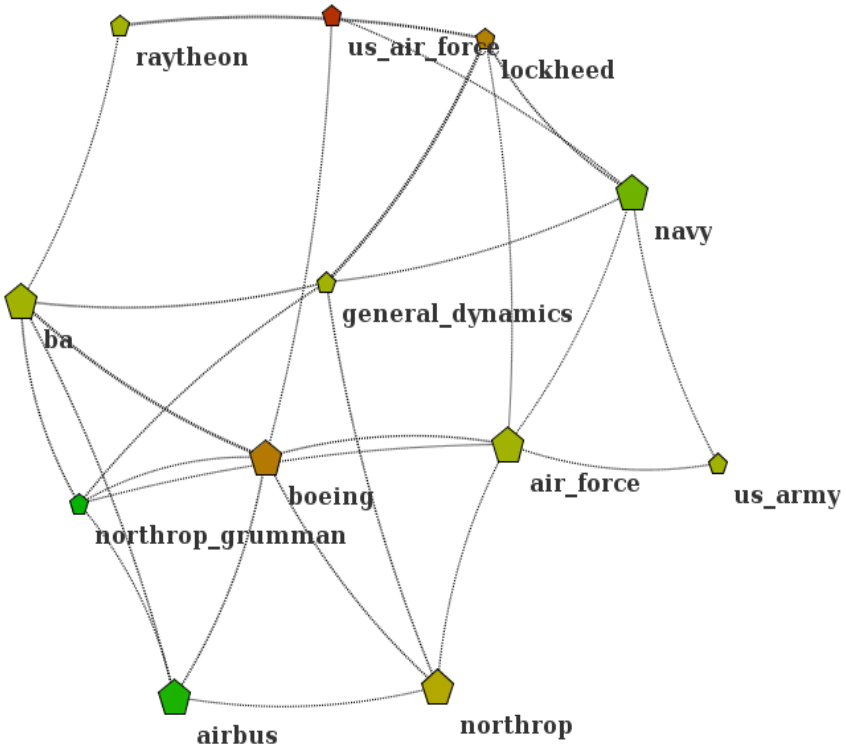

Figure 7.3.4: US organisation network restricted to the second largest cluster; September 2008

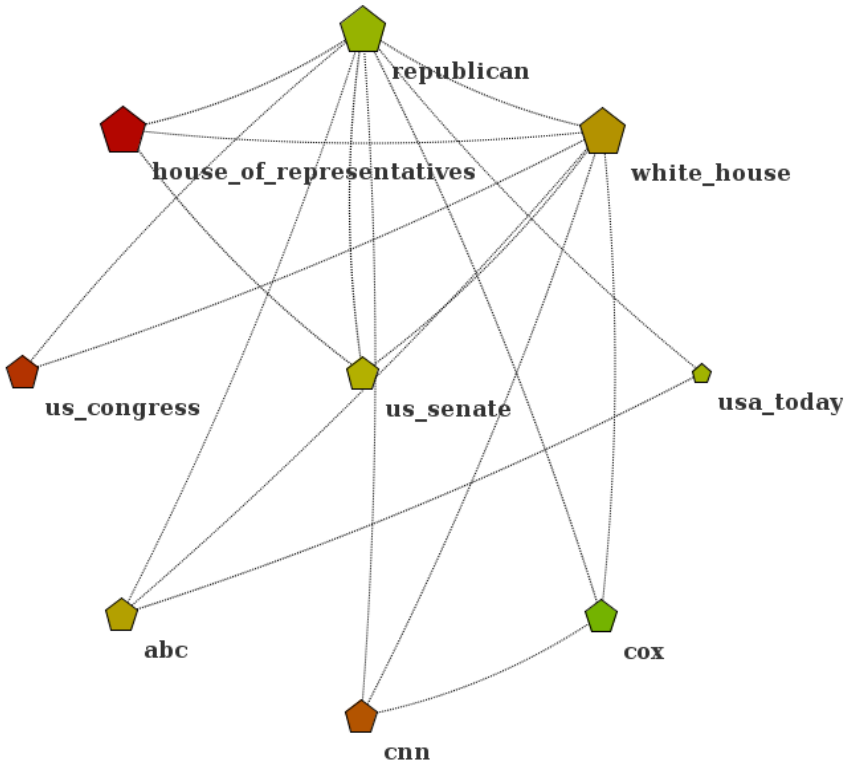

Figure 7.3.5: US organisation network restricted to the third largest cluster; September 2008

8

Assessment, Conclusion and Future Work

This thesis has examined the hypothesis that CNT can be made operational and aspects of it empirically tested via a combination of text analysis, network analysis and machine learning techniques. One hypothesis of CNT that was tested in the first experiment is the potential predictive relationship between confidence (as defined by the emotional conviction of a social group, such as the participants of an economy) and economic growth. A further hypothesis of CNT that has been tested is the potential to use the derived methodology to detect the phenomena known as *divided state* mentality with respect to a given conviction narrative, represented by a disconnect between conviction and reality for a given social group (where the group dynamic is referred to as *groupfeel*). The final hypothesis that has been tested states that those narratives that emerge as *dominant* narratives in a social group (e.g., an economy) have a predominance of either approach or avoidance emotion and that the periods when dominant narratives emerge therefore have a negative impact on the stability of the system (e.g., financial stability).

This chapter assesses the contributions or otherwise made through the design and tests of the three experiments as well as the construction of the RSS methodology and its evaluation in Chapter 4. This is followed by some concluding remarks and a discussion on future work.

8.1 Assessment

The thesis set out to derive a methodology to test some of the main hypotheses postulated by CNT. The two main hypotheses postulated by CNT are formulated as follows.

1. It is hypothesised that a relative increase in the dominance of approach or avoidance emotions (loosely defined as *excitement* and *anxiety*) in overall economic narratives (i.e., from a macro point of view, in a country, sector, region) indicates a change in economic confidence and so actions that impact the economy and economic growth.

2. It is hypothesised that a significant relative gap in the dominance of excitement over anxiety around particular topics indicate that thinking and decision-making around these topics have become unbalanced, creating a "divided state" - suggesting the build-up of risk.

A further hypothesis is postulated in this thesis as follows.

3. It is hypothesised that a significant shift in macro confidence (i.e., the macro RSS series) occurs exactly when a dominant narrative has emerged. In other words, the connection between the macro and the micro is such that a micro story becomes dominant whenever there is a large shift in the macro series. Furthermore, the formation of such a dominant narrative is detrimental to financial stability, as financial markets rely on heterogeneity of views.

The experiments conducted to test the above hypotheses lead to several scientific contributions, listed below.

8.1.1 Contribution 1: A Methodology for Lexicon Evaluation in Time Series Sentiment Analysis

Chapter 4 derives a methodology to measure the relative balance between approach (excitement) and avoidance (anxiety) emotions. Having first argued for a word-lexicon approach as opposed to a machine-learning classification approach to sentiment analysis for the purpose of operationalising CNT, several criteria to measure the accuracy of a given lexicon are defined. This facilitates the evaluation of alternative lexicons for the purpose of capturing two latent emotional time series variables. It is found that the purpose-built lists of excitement and anxiety words perform well, with the anxiety list outperforming all other lists on most criterion. In particular, words that potentially refer to economic concepts, such as 'boost', are found to be somewhat

orthogonal (in the sense of being uncorrelated) to generic emotion words. This supports a key insight about word-list based approaches to sentiment analysis, namely that the lists of emotion words (features) should ideally exclude context specific words, if the assumption is that all words in a given list should tend to correlate with the 'target' emotional signal. The methodology for evaluating this property of a given lexicon and for adjusting the lexicon based on these criteria is the first major contribution of the thesis.

8.1.2 Contribution 2: Predicting GDP and MCI using RSS

Chapter 5 (experiment one) applies the RSS methodology developed in Chapter 4 to test the first hypothesis listed above, if changes in economic confidence impact the economy. It is found that the confidence index derived for the US economy through an analysis of Reuters news articles published in New York and Washington can be used as a significant explanatory variable in a regression of US GPD growth, even when controlling for other forward-looking variables such as asset prices (and other measures of confidence and anxiety such as the Michigan Consumer Sentiment Index and the VIX, which are in fact shown to be Granger-caused by RSS indices extracted from other data sources and/or Reuters[1]). This provides significant support for the first hypothesis. Furthermore, it shows the potential to forecast, as well as nowcast[2], GDP growth using the RSS methodology. This is the second major contribution of the thesis.

8.1.3 Contribution 3: Operationalising Phantastic Objects, Divided States & Groupfeel using RSS

Chapter 6 (experiment two) applies the RSS methodology to particular narratives (topics) hypothesised to show characteristics of phantastic object narratives. It is found that stories about Fannie Mae exhibited a significant decline in anxiety relative to excitement over the period 2005 through 2007, despite the fact that housing prices started to decline almost 18 months before the end of this trend in RSS. The pattern is consistent with expectations from CNT regarding phantastic object narratives. Similar relative sentiment patterns are observed when the analysis is focused on Enron emails regarding two specific business projects. The email datasource allowed for the construction of a social network showing the communication channels between

[1] A RSS index extracted from broker reports, a datasource more likely to contain subjective beliefs about the state of the economy, can be used to make substantially better predictions of the Michigan index than those made by economists surveyed by Reuters [78]. A further index extracted from market comments produced at the Bank of England is shown to weakly Granger-cause the VIX.

[2] The methodology was used in [77] to nowcast (predict the current value of) US GDP growth.

Enron employees. By applying social network analysis techniques, in particular graph clustering, it is shown that a key social group experienced significantly different sentiment patterns concerning these two business projects relative to the rest of the network. The group became substantially more anxious in the second half of the period; evidence of a divided state collapsing within this particular group. The experiment provides support for the second hypothesis. It also suggests that it is possible to distinguish the emotional patterns between different social groups, and that social context may therefore play a significant role in determining an individual agent's emotional state. In other words, supporting the idea that text analysis can be used to measure groupfeel-type phenomena. This is the third major contribution of the thesis.

8.1.4 Contribution 4: A Methodology to Measure DNF and Establishing its Relationship with RSS and Financial Stability

Chapter 7 (experiment three) develops a new methodology to relate entities within narratives to one another and the overall structure of the narrative landscape to the economic confidence index of the US created in Chapter 4. This methodology is used to test the third hypothesis. It is shown that shifts in the economic confidence index is correlated with the emergence of a dominant organisation cluster. By application of Granger-causality tests it is further shown that new indices of *dominant narrative formation* developed as part of the third experiment carry predictive information relating to financial stability, in the form of the VIX and the S&P 500 index. Thus, the results provide support for the third hypothesis. The particular narratives do not necessarily have to be determined in order to derive useful indicators of risks to financial stability, it may in some cases be enough to determine the overall structure of the narrative landscape and how that structure evolves over time. This is the fourth and final major contribution of the thesis.

8.2 Conclusion

Conviction narrative theory (CNT) is a theory of decision-making that asserts that, when faced with uncertainty, agents are able to act by constructing narratives that yield emotional conviction. This thesis sets out to develop a methodology to make CNT operational and empirically testable via a combination of novel text analysis, network analysis and machine leaning techniques.

The methodology is directed by the theory, thus limiting problems of spurious correlations. This is a central theme emphasised throughout the thesis. Because the true model of any system is inherently uncertain, e.g., the true model of the financial

system and the economy is unknown and perhaps even unknowable, predictions made from simple (parsimonious) and theoretically justified models are more likely to remain accurate in the presence of new data than are predictions made using models with more parameters and interactions. This general principle, or heuristic, is often referred to as *Occam's razor*. It is argued, in the context of a generic discussion, that a methodology directed by theory is effectively biased (both the model space and the features considered relevant by the theory) and that this introduced bias gives confidence that predictions made will continue to be valid for a substantially longer period of time as new data arrives, than are theory-agnostic predictions. The bias-variance tradeoff provides a statistical perspective on the issue, as does Bayesian statistics; as a theory can be considered a strong prior belief regarding what features should be considered important.

Through the construction of the RSS methodology in Chapter 4 and the three main experiments in Chapters 5, 6 and 7 the theory is made operational in a highly parsimonious manner. The key concepts of conviction (in the form of the relative balance between approach and avoidance emotions), divided state (the disconnect between conviction and reality), groupfeel (socially dependent emotional state) and phantastic object (a conviction narrative subject to divided state mentality and, on the macro level, also groupfeel) are explored and tested on available datasources and case studies. Support for CNT is found in all cases explored by the thesis. Both the method used to evaluate, and potentially adjust, a given emotion lexicon used to construct a confidence index and the confidence index itself will likely prove useful both for academic research on the effect of sentiment on financial and economic systems and for policy-makers interested in using the techniques to monitor and forecast important policy-relevant variables in a more practical setting. The third and final experiment helps to qualitatively explain large shifts in the macro RSS series, the confidence index, by showing that such shifts take place when dominant narratives have emerged. The methodology derived through the experiment may also prove useful in future research attempting to warn about emerging narratives showing phantastic-object characteristics.

The objectives set out in Chapter 1.2 have been supported by the analysis developed throughout the thesis. However, this is a new area of research and much remains to be explored, but the methodology and findings developed and presented in this thesis clearly show the potential of analysis of unstructured data to shed light on socio-psychological and economic theory. As such, the thesis has been successful in what it initially set out to achieve and serves as a foundation for further research in the area.

8.3 Future Work

Several areas of research further developing the RSS and DNF methodologies could be pursued that would build on the work of this thesis. This section summarises some of the possible extensions into several related future potential work streams.

8.3.1 Future Work Stream 1: Emotion Lexicons

Future research might beneficially extend the emotion lexicons making use of principal component analysis to test whether or not to include words if they correlate significantly with the principal component. However, it should be stressed that a theoretical focus should be maintained, implying social-psychological judgement and expertise should be exercised. Such an empirical test would likely benefit from more and varied datasources. In this thesis the only datasource used to test the signal measured by a wordlist was the US Reuters news archive. A simple method to find candidate words to include in a list might be to correlate the word frequency series of all words in a corpus with the principal component of a list and test for inclusion. This method relies on the assumption that the original list of words is fairly clear and consistent, and that the principal component signal of the list is close to the intended target signal. Given the concerns about spurious correlations raised in Chapter 2.1, whatever method is used to refine a list should be augmented with judgement and possibly also survey data.

8.3.2 Future Work Stream 2: RSS Methodology

The RSS methodology itself could potentially be improved by moving beyond a single-word approach into n-grams or more complex natural-language-processing methods. However, the analysis of Chapter 4 makes clear some of the difficulties with the simpler bag-of-words unigram approach. A further potential improvement, not discussed thus far, relates to how the daily emotional scores could be aggregated to longer periods. All experiments aggregate daily observations equally into a monthly or quarterly series, but other strategies, such as exponential smoothing, might prove more effective strategies to create leading indicators. Smoothing techniques that give more weight to more recent documents might be particularly successful if attempting to predict time series variables representing 'spot-prices' or other more 'instantaneous' variables. Given the findings on the robustness of different ways to aggregate the emotion words, e.g., adjusting for individual word frequencies, it might also prove fruitful to pursue further tests along these lines.

8.3.3 Future Work Stream 3: Topic Identification

The US macro confidence index can likely be improved through a more careful selection of which articles to analyse. The heuristic of selecting articles from the New York and Washington offices is unlikely the most optimal selection procedure. The method to filter documents based on a given topic can potentially be improved beyond a regular expression match, potentially using semantic models like LSA or topic models like LDA, in combination with machine-learning classifiers. The experiment assumes that words within a certain proximity to a pattern are related to that pattern, the method relying on this fairly strong assumption could potentially be refined. However, pattern matching is both a computationally efficient and transparent way to filter documents. Indeed, the news analysis system developed by the European Commission as mentioned in Chapter 2.1 is based on exactly the same ideas and has also been shown to be very effective. Another means by which to identify topics to analyse using the RSS methodology could be to dynamically cluster the entity networks of Chapter 7.

8.3.4 Future Work Stream 4: Social and Cognitive Risk Detection

Combining the second and third experiments, and potentially several other techniques (e.g., LSA and clustering), to understand how phantastic object narratives are formed and evolve, via social networks and the narrative entity network, and to be able to quantify them at an early stage is a major future research goal. The network data model used in Chapter 7 can easily be extended to include information about who is talking about what and with whom, by simply encoding this information into the graph database. The database can then be queried efficiently both in terms of the social network and the narrative network to achieve very powerful analysis.

8.3.5 Future Work Stream 5: Further Predictions and Improved Predictive Models

Very simple and straightforward tests for statistical relationships (e.g., linear regression and Granger-causality) between time series variables have been used in all experiments. Although such simple models provide increased confidence in any relationships found, it is likely that other more elaborate techniques are superior for the purpose of improving the accuracy and strength of the relationships. More thorough out-of-sample predictions could be made of those variables found to be Granger-caused by the RSS and DNF measures. In particular, further research might beneficially improve on predictions of the Michigan Consumer Sentiment index by updating the BROKER RSS series and by a more careful selection of which reports to

analyse. More formal forecasting could potentially be carried out and compared to alternative forecasts. The other work streams can potentially be calibrated by this fifth stream.

References

[1] May 2015. URL https://gate.ac.uk/.

[2] May 2015. URL http://www.wjh.harvard.edu/~inquirer/.

[3] May 2015. URL http://igraph.org/redirect.html.

[4] May 2015. URL http://jung.sourceforge.net/.

[5] May 2015. URL http://neo4j.com/.

[6] May 2015. URL http://thomsonreuters.com/en/products-services/financial/news-and-insight/machine-readable-news.html.

[7] May 2015. URL http://sentiwordnet.isti.cnr.it/.

[8] May 2015. URL https://uk.finance.yahoo.com/.

[9] H. Abdi. *Encyclopedia of Measurement and Statistics*, chapter Bonferroni and Šidák corrections for multiple comparisons. Sage, 2007.

[10] Peter Abell. Some aspects of narrative method. *Journal of Mathematical Sociology*, 18(2):93–134, 1993.

[11] J. Adibi and J. Shetty. Enron dataset. URL http://www.isi.edu/ adibi/Enron/Enron.htm.

[12] David Aikman, Mirta Galesic, Gerd Gigerenzer, Sujit Kapadia, Konstantinos Katsikopoulos, Amit Kothiyal, Emma Murphy, and Tobias Neumann. Taking uncertainty seriously: simplicity versus complexity in financial regulation. *Bank of England Financial Stability Paper*, May 2014.

[13] H. Akaike. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6):716–723, 1974.

[14] R F Baumeister and E J Masicampo. Conscious thought is for facilitating social and cultural interactions: how mental simulations serve the animal-culture interface. *Psychological Review*, 117(4), 2010.

[15] Benjamini and Hochberg. Controlling the false discovery rate: A practival and pwoerful approach to multiple testing. *Journal of the Royal Statistical Society, Serie B*, 57:289–300, 1995.

[16] Steven Bird, Edward Loper, and Ewan Klein. *Natural Language Processing with Python*. O'Reilly Media Inc, May 2009. URL `http://www.nltk.org/`.

[17] S. Blair-Goldensohn, K. Hannan, R. McDonald, T. Neylon, G. A. Reis, and J. Reynar. Building a sentiment summarizer for local service reviews. In *NLP Challenges in the Information Explosion Era (NLPIX)*, 2008.

[18] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, pages 993–1022, 2003.

[19] M. M. Bradley and P. J. Lang. Affective norms for english words (anew): Instruction manual and affective ratings. Technical report c-1, The Center for Research in Psychophysiology, University of Florida, 1999.

[20] Signe Bray, Shinsuke Shimojo, and John P. O'Doherty. Human medial orbitofrontal cortex is recruited during experience of imagined and real rewards. *Journal of Neurophysiology*, 103(5):2506–2512, May 2010.

[21] T. S. Breusch. Testing against general autoregressive and moving average error models when the regressors include lagged dependent variables. *Econometrica*, 46:1293–1302, 1978.

[22] T. S. Breusch and A. R. Pagan. A simple test for heteroscedasticity and random coefficient variation. *Econometrica*, 47(5):1287–1294, 1977.

[23] Eric Brill. A simple rule-based part of speech tagger. In *Proceedings of the third conference on Applied natural language processing (ANLC '92)*, pages 152–155, Stroudsburg, PA, USA, 1992. Association for Computational Linguistics.

[24] Jerome Bruner. The narrative construction of reality. *Critical Inquiry*, 18(1-21), 1991.

[25] Hyunyoung Choi and Hal Varian. Predicting the present with google trends. *Economic Record*, 88:2–9, June 2012.

[26] Kimberly Chong and David Tuckett. Constructing conviction through action and narrative: how money managers manage uncertainty and the consequence for financial market functioning. *Socio-Economic Review*, pages 1–26, May 2014.

[27] Richard Colbaugh and Kristin Glass. Leveraging sociological models for prediction i: Inferring adversarial relationships. In *IEEE International Conference on Intelligence and Security Informatics*, pages 66–71. IEEE, 2012.

[28] Richard Colbaugh and Kristin Glass. Leveraging sociological models for prediction ii: Early warning for complex contagions. In *IEEE International Conference on Intelligence and Security Informatics*, pages 72–77. IEEE, 2012.

[29] S. Deerwester, S. T. Dumas, G. W. Furnas, T. K. Landauer, and R. Harshman. Indexing by latent semantic analysis. *Journal of the society for information science*, 41(6), 1990.

[30] Victor DeMiguel, Lorenzo Garlappi, and Raman Uppal. Optimal versus naive diversification: How inefficient is the 1/n portfolio strategy? *Review of Financial Studies*, 22:1915–1953, 2009.

[31] K. Dickersin, S. Chan, T. C. Chalmers, H. S. Sacks, and H. Smith Jr. Publication bias and clinical trials. *Controlled Clinical Trials*, 8:343–353, 1987.

[32] J. Diesner and K. M. Carley. Exploration of communication networks from the enron email corpus. In *Proceedings of Workshop on Link Analysis, Counterterrorism and Security*, pages 3–14. SIAM International Conference on Data Mining, 2005.

[33] J. Diesner, K. M. Carley, and L. Tambayong. Extracting socio-cultural networks of the sudan from open-source, large-scale text data. *Computational & Mathematical Organization Theory*, 18:328–339, 2012.

[34] Peter Sheridan Dodds, Eric M Clark, Suma Desu, Morgan R Frank, Andrew J Reagan, Jake Ryland Williams, Lewis Mitchell, Kameron Decker Harris, Isabel M Kloumann, James P Bagrow, Karine Megerdoomian, Matthew T McMahon, Brian F Tivnan, and Christopher M Danforth. Human language reveals a universal positivity bias. *Proceedings of the National Academy of Sciences*, 112(8):2389–2394, 2014.

[35] Ruth Dunn, Frank Dudbridge, and Christopher M. Sanderson. The use of edge-betweenness clustering to investigate biological function in protein interaction networks. *BMC Bioinformatics*, 6(39), March 2005.

[36] J. Durbin. *1. Distribution Theory for Tests Based on the Sample Distribution Function*, chapter 1, pages 1–64. doi: 10.1137/1.9781611970586.ch1. URL `http://epubs.siam.org/doi/abs/10.1137/1.9781611970586.ch1`.

[37] J. Durbin and G. S. Watson. Testing for serial correlation in least squares regression i. *Biometrica*, 37:409–428, 1950.

[38] J. Durbin and G. S. Watson. Testing for serial correlation in least squares regression ii. *Biometrica*, 38:159–178, 1951.

[39] J. Durbin and G. S. Watson. Testing for serial correlation in least squares regression iii. *Biometrica*, 58:1–19, 1971.

[40] Philippa K. Easterbrook, Jesse A. Berlin, Ramana Gopalan, and David R. Matthews. Publication bias in clinical research. *The Lancet*, 337:867–872, April 1991.

[41] B. Efron. Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, 7(1):1–26, 1979.

[42] Milton Friedman. *Essays in Positive Economics*. 1953.

[43] Michelle Girvan and M. E. J. Newman. Community structure in social and biological networks. Sfi working paper, Santa Fe, 2001.

[44] C. W. J. Granger. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica*, 37(3):424–438, 1969.

[45] C. W. J. Granger and P. Newbold. Spurious regressions in econometrics. *Journal of Econometrics*, 2:111–120, 1974.

[46] L. A. Green and D. R. Mehr. What alters physicians' decision to admit to the coronary care unit? *Journal of Family Practice*, 45:219–226, 1997.

[47] Till Grüne-Yanoff. Paradoxes of rational choice theory. *Handbook of Risk Theory*, pages 499–516, 2012.

[48] Stephen Hansen, Michael McMahon, and Andrea Prat. Transparency and Deliberation within the FOMC: A Computational Linguistics Approach. CEP Discussion Papers dp1276, Centre for Economic Performance, LSE, June 2014. URL `http://ideas.repec.org/p/cep/cepdps/dp1276.html`.

[49] Frank E. Jr. Harrell. *Regression modeling strategies*. Springer-Verlag, New York, 2001.

[50] Zonglu He and Koichi Maekawa. On spurious granger causality. *Economics Letters*, 73:307–313, 2001.

[51] P. M. Healy and K. G. Palepu. The fall of enron. *Journal of Eonomic Perspectives*, 17(2):3–26, 2003.

[52] M. Hepple. Independence and commitment: Assumptions for rapid training and execution of rule-based pos taggers. In *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics (ACL-2000)*, Hong Kong, October 2000.

[53] Irving L. Janis. *Groupthink: Psychological Studies of Policy Decisions and Fiascoes*. Cengage Learning, 2 edition, May 1982.

[54] Daniel Kahneman and Dan Lovallo. Timid choices and bold foreacasts: A cognitive perspective on risk taking. *Management Science*, 39(1):17–31, 1993.

[55] Norbert L. Kerr. Harking: Hypothesizing after the results are known. *Personality and Social Psychology Review*, 2(3):196–217, 1998.

[56] B. Klimt and Y. Yang. Introducing the enron corpus. In *First Conference on Email and Anti-Spam (CEAS)*, Mountain View, CA, 2014.

[57] B. Klimt and Y. Yang. The enron corpus: a new dataset for email classification research. In *European Conference on Machine Learning (ECML)*, pages 217–226, Pisa, Italy, 2014.

[58] Michal Kosinski, David Stillwell, and Thore Graepel. Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the National Academy of Sciences*, 110(15):5802–5805, 2013.

[59] D. Kwiatkowski, P. C. B. Phillips, P. Schmidt, and Y. Shin. Testing the null hypothesis of stationarity against the alternative of a unit root. *Journal of Econometrics*, 54:159–178, 1992.

[60] P. J. Lang and M. M. Bradley nad B. N. Cuthbert. Emotion and motivation: measuring affective perception. *Journal of Clinical Neurphysiology*, 15(5): 397–408, 1998.

[61] Benjamin Letham, Cynthia Rudin, Tyler H Mccormick, and David Madigan. Interpretable classifiers using rules and bayesian analysis: Building a better stroke prediction model. 2013.

[62] M Lienou, H Maitre, and M Datcu. Semantic annotation of satellite images using latent dirichlet allocation. *Geoscience and Remote Sensing Letters, IEEE*, 7(1):28–32, 2009.

[63] Mingfeng Lin, Henry C. Lucas, and Galit Shmueli. Too big to fail: Large samples and the p-value problem. *Articles in Advance*, pages 1–12, 2013.

[64] Marielle Linting, Bart Jan Van Os, and Jacqueline J. Meulman. Statistical significance of the contribution of variables to the pca solution: An alternative permutation strategy. *Psychometrica*, 76(3):440–460, July 2011.

[65] Tim Loughran and Bill McDonald. When is a liability not a liability? textual analysis, dictionaries, and 10-ks. *The Journal of Finance*, 66(1):35–65, February 2011.

[66] Robert Jr Lucas. *Econometric policy evaluation: A critique*, volume 1, pages 19–46. January 1976. URL `http://ideas.repec.org/a/eee/crcspp/v1y1976ip19-46.html`.

[67] C. D. Manning, P. Raghavan, and H. Schütze. *Introduction to Information Retrieval*. Cambridge University Press, 2008. URL `http://nlp.stanford.edu/IR-book/html/htmledition/irbook.html`.

[68] Raymond A. Mar and Keith Oatley. The function of fiction is the abstraction and simulation of social experience. *Perspectives on Psychological Science*, 3(3): 173–192, 2008.

[69] Girish Maskeri, Santonu Sarkar, and Kenneth Heafield. Mining business topics in source code using latent dirichlet allocation. In *Proceedings of the 1st India Software Engineering Conference*, ISEC '08, pages 113–120, New York, NY, USA, 2008. ACM.

[70] Viktor Mayer-Schonberger and Kenneth Cukier. *Big Data: A Revolution That Will Transform How We Live, Work and Think*. John Murray, October 2013.

[71] B. McLean and P. Elkind. *The Smartest Guys in the Room: The Amazing Rise and Scandalous Fall of Enron*. Portfolio Trade, 2004.

[72] Madan Lal Mehta. *Random Matrices*, volume 142 of *Pure and Applied Mathematics*. Academic Press, November 2000.

[73] N. Metropolis and S. Ulam. The monte carlo method. *Journal of American Statistical Association*, 1949.

[74] Kevin P. Murphy. *Machine Learning: A Probabilistic Perspective*. MIT Press, Cambridge Masachusetts, London England, 2012.

[75] M. E. J. Newman. The structure and function of complex networks. *aps.arxiv.org/abs/cond-mat/0303516*, March 2003.

[76] M. E. J. Newman. Finding community structure in networks using the eigenvectors of matrices. *Physics Review E 74 036104*, 2006.

[77] Rickard Nyman, David Gregory, Sujit Kapadia, Paul Ormerod, Robert Smith, and David Tuckett. News and narratives in financial systems: exploiting big data for systemic risk assessment. Washington, December 2014. Federal Reserve Bank of Cleveland and Office of Financial Research.

[78] Rickard Nyman, Paul Ormerod, Robert Smith, and David Tuckett. Big data and economic forecasting: A top-down approach using directed algorithmic text analysis. Frankfurt, April 2014. European Central Bank.

[79] Rickard Nyman, Paul Ormerod, and Robert Smith. Measuring stability, clarity, and consistency of word frequency sentiment analysis. June 2015.

[80] Paul Ormerod and Craig Mounfield. Random matrix theory and the failure of macro-economic forecasts. *Physica A: Statistical Mechanics and its Applications*, 280(3):497–504, 2000.

[81] Paul Ormerod, Rickard Nyman, and Alexander Bentley. On promises and pitfalls of big data: A case study of google flu trends. European Commission, Brussels, October 2014. Global Systems Science.

[82] B. Pang and L. Lee. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In *Proceedings of the ACL*, pages 115–124, 2005.

[83] B. Pang, L. Lee, and S. Vaithyanathan. Thumbs up? sentiment classification using machine learning techniques. In *Proceedings of EMNLP*, pages 79–86, 2002.

[84] Bo Pang and Lillian Lee. A sentimental education: Sentiment analysis using subjectivity. In *Proceedings of ACL*, pages 271–278, 2004.

[85] Bo Pang and Lillian Lee. Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 2(1-2):1–135, January 2008.

[86] Dan Pelleg and Andrew W. Moore. X-means: Extending k-means with efficient estimation of the number of clusters. In *Proceedings of the Seventeenth International Conference on Machine Learning*, ICML '00, pages 727–734, San Francisco, CA, USA, 2000. Morgan Kaufmann Publishers Inc. ISBN 1-55860-707-2. URL `http://dl.acm.org/citation.cfm?id=645529.657808`.

[87] W. Ploberger and W Krämer. The cusum test with ols residuals. *Econometrica*, 60:271–285, 1992.

[88] J. B. Ramsey. Tests for specification error in classical linear least squares regression analysis. *Journal of the Royal Statistical Society, Series B*, 31: 350–371, 1969.

[89] Robert Rosenthal. The "file drawer problem" and tolerance for null results. *Psychological Bulletin*, 86(3):638–641, 1979.

[90] Magnus Sahlgren. An introduction to random indexing. In *In Methods and Applications of Semantic Indexing Workshop at the 7th International Conference on Terminology and Knowledge Engineering, TKE 2005*, 2005.

[91] S. E. Said and D. A. Dickey. Testing for unit roots in autoregressive-moving average models of unknown order. *Biometrica*, 71:599–607, 1984.

[92] Marcel Salathé and Shashank Khandelwal. Assessing vaccination sentiments with online social media: Implications for infectious disease dynamics and control. *PLoS Computational Biology*, 7(10), 2011.

[93] G. Salton, A. Wong, and C. S. Yang. A vector space model for automatic indexing. *Communications of the ACM*, 18(11):613–620, November 1975.

[94] Steven L. Scott and Hal R. Varian. Bayesian variable selection for nowcasting economic time series. *National Bureau of Economic Research Working Paper*, (19567), October 2013.

[95] A. M. Sengupta and P. P. Mitra. Distributions of singular values for some random matrices. *arXiv:cond-mat/9709283*, September 1997.

[96] Claude E. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27(3):379–423, 1948.

[97] Galit Shmueli. To explain or to predict? *Statistical Science*, 25(3):289–310, 2010.

[98] Hans-Werner Sinn. *Casino Capitalism. How the Financial Crisis Came About and What Needs to Be Done Now*. Oxford University Press, Oxford, 2010.

[99] Adam Smith. *An Inquiry into the Nature and Causes of the Wealth of Nations*. W. Strahan and T. Cadell, London, 1776.

[100] Gordon K. Smyth and Belinda Phipson. Permutation p-values should never be zero: Calculating exact p-values when permutations are randomly drawn. *Statistical Applications in Genetics and Molecular Biology*, 9(1):Article 39, October 2010.

[101] B. Snyder and R. Barzilay. Multiple aspect ranking using the good grief algorithm. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics,*, pages 300–307, 2007.

[102] Cindy K. Soo. Quantifying animal spirits: News media and sentiment in the housing market. *Ross School of Business Paper No. 1200*, August 2013.

[103] Viktor M. Strauss. Emotional values of words in finance: Anxiety about losses and excitement about gains. M.sc. thesis in social cognition, University College London, Gower Street, 2013.

[104] Saatviga Sudhahar, Guiseppe A. Veltri, and Nello Cristianini. Automated analysis of the us presidential elections using big data and network analysis. *Big Data & Society*, 2(1), March 2015.

[105] Chenhao Tan, Long Jiang, Lillian Lee, Ming Zhou, Jie Tang, and Ping Li. User-level sentiment analysis incorporating social networks. In *In KDD*, pages 1397–1405, 2011.

[106] P. C. Tetlock, M. Saar-Tsechansky, and S. Macskassy. More than words: Quantifying language to measure firms' fundamentals. *The Journal of Finance*, 63(3):1437–1467, 2008.

[107] Paul C. Tetlock. Giving content to investor sentiment: The role of media in the stock market. *The Journal of Finance*, 62(3):1139–1168, 2007.

[108] Paul C. Tetlock. All the news that's fit to reprint: Do investors react to stale information? *Review of Financial Studies*, 24:1481–1512, 2011.

[109] Hiro T. Toda and Taku Yamamoto. Statistical inference in vector autoregression with possibly integrated processes. *Journal of Econometrics*, 66:225–250, 1995.

[110] Jean-Claude Trichet. Reflections on the nature of monetary policy non-standard measures and finance theory. European Central Bank, November 2010.

[111] David Tuckett. *Minding the markets: An Emotional Finance View of Financial Instability*. Palgrave Macmillan, May 2011.

[112] David Tuckett and Milena Nikolic. The role of conviction and narrative in decision-making under deep uncertainty. UCL Centre for the Study of Decision-Making Uncertainty, 2015.

[113] David Tuckett and Richard Taffler. Phantastic objects and the financial market's sense of reality: A psychoanalytic contribution to the understanding of stock market instability. *The International Journal of Psychoanalysis*, 89: 389–412, 2008.

[114] David Tuckett, Robert Smith, and Rickard Nyman. Tracking phantastic objects: A computer algorithmic investigation of narrative evolution in unstructured data sources. *Social Networks*, 38:121–133, 2014.

[115] David Tuckett, Paul Ormerod, Rickard Nyman, and Robert Smith. Improving economic prediction: A new method for measuring economic confidence and its impact on the evolution of the us economy. *In Press*, 2015.

[116] P. D. Turney. Thumbs up or thumbs down? semantic orientation applied to unsupervised classification of reviews. In *ACL '02 Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pages 417–424, 2002.

[117] Alfredo Vellido, José D Martín-Guerrero, and Paulo Lisboa. Making machine learning models interpretable. In *20th European Symposium on Artificial Neural Networks*, 2012.

[118] A. B. Warriner, V. Kuperman, and M. Brysbaert. Norms of valence, arousal, and dominance for 13,915 english lemmas. *Behavioural Research Methods*, pages 1–17, 2013.

[119] Duncan J Watts. Common sense and sociological explanations. *American Journal of Sociology*, 120(2):313–351, 2014.