

© 2014

John Robert Yaros

ALL RIGHTS RESERVED

# DATA MINING PERSPECTIVES ON EQUITY SIMILARITY PREDICTION

By

JOHN ROBERT YAROS

A dissertation submitted to the  
Graduate School—New Brunswick  
Rutgers, The State University of New Jersey  
in partial fulfillment of the requirements  
for the degree of  
Doctor of Philosophy  
Graduate Program in Computer Science  
written under the direction of  
Tomasz Imieliński  
and approved by

---

---

---

---

New Brunswick, New Jersey

October, 2014

## **ABSTRACT OF THE DISSERTATION**

# **Data Mining Perspectives on Equity Similarity Prediction**

**By JOHN ROBERT YAROS**

**Dissertation Director:**

**Tomasz Imieliński**

Accurate identification of similar companies is invaluable to the financial and investing communities. To perform relative valuation, a key step is identifying a “peer group” containing the most similar companies. To hedge a stock portfolio, best results are often achieved by selling short a hedge portfolio with future time series of returns most similar to the original portfolio - generally those with the most similar companies. To achieve diversification, a common approach is to avoid portfolios containing any stocks that are highly similar to other stocks in the same portfolio.

Yet, the identification of similar companies is often left to hands of single experts who devise sector/industry taxonomies or other structures to represent and quantify similarity. Little attention (at least in the public domain) has been given to the potential that may lie in data-mining techniques. In fact, much existing research considers sector/industry taxonomies to be ground truth and quantifies results of clustering algorithms by their agreement with the taxonomies.

This dissertation takes an alternate view that proper identification of relevant features and proper application of machine learning and data mining techniques can achieve results that rival or even exceed the expert approaches. Two representations of similarity are considered: 1) a pairwise approach, wherein a value is computed to quantify the similarity

for each pair of companies, and 2) a partition approach analogous to sector/industry taxonomies, wherein the universe of stocks is split into distinct groups such that the companies within each group are highly related to each other. To generate results for each representation, we consider three main datasets: historical stock-return correlation, equity-analyst coverage and news article co-occurrences. The latter two have hardly been considered previously. New algorithmic techniques are devised that operate on these datasets. In particular, a hypergraph partitioning algorithm is designed for imbalanced datasets, with implications beyond company similarity prediction, especially in consensus clustering.

## Preface

Portions of this dissertation draw from research previously published (Yaros and Imieliński, 2013a,b, 2014a,c,b).

## Acknowledgments

Without the kind support of many people, I would not have been able to complete my doctoral studies. First, I thank my advisor, Tomasz Imieliński, who has spent countless hours with me in his office, the Busch campus center and many coffee shops throughout New Jersey. I truly appreciate his patience in the face of my habitual skepticism and have enjoyed many of our discussions, including those that have deviated from research to humor, politics, Polish history and life in general. As his teaching assistant, I was able to participate in many of his unconventional but very effective teaching methods, and believe the students received valuable training, although none were as fortunate as me to receive his greater attention as doctoral student.

S. Muthukrishnan also receives my humble gratitude. Through his Algorithms II class and by including me in his research projects, I became immersed in the domain of advertising — a very different world than my usual domain of finance. His expertise offers insights that few others can provide. I am also grateful for his constructive feedback as a committee member.

I also thank the other members of my committee, Vladimir Pavlovic and Walter Tackett. Through his Intro. to A.I. and Machine Learning courses, Prof. Pavlovic provided much of the foundation for my thesis studies. Walter Tackett help me in finding a job in addition to providing insightful and knowledgeable feedback on this thesis. Liviu Iftode also provided valuable feedback during my qualification exam.

In addition to faculty, much of my learning came from fellow students. In particular, Michael Wunder was much an older brother to me and greatly helped me to start research. He included me in his multi-agent projects and also provided valuable advice on managing the journey through graduate school. I also thank Qiang Ma, Darja Krushevskaja, Edan Harel and Vasisht Goplan for enabling me to quickly participate in the advertising research

lead by Muthu.

Finally, I am much indebted to my family for their years of support. My wife, Katarina, was often the first sounding board for ideas as well as my “ace in the hole” for any mathematical challenges. I appreciate her tolerating a meager salary for so many years. I am also very appreciative of the support of my parents, who took an early retirement and moved from New Mexico to New Jersey in order to help with our children as I studied. Having children in graduate school would have otherwise been impossible.

## Dedication

For Kika and Robbie



# Table of Contents

<b>Abstract</b> . . . . .	ii
<b>Preface</b> . . . . .	iv
<b>Acknowledgments</b> . . . . .	v
<b>Dedication</b> . . . . .	vii
<b>1. Introduction</b> . . . . .	1
1.1. Contributions . . . . .	4
1.1.1. New Perspectives . . . . .	4
1.1.2. Methods and Tools . . . . .	5
1.1.3. Applications . . . . .	6
1.2. Dissertation Structure . . . . .	7
<b>2. Background &amp; Related Work</b> . . . . .	8
2.1. Correlation as a Measure of Similarity . . . . .	8
2.1.1. Possible Alternatives . . . . .	11
Cointegration . . . . .	11
Financial Statement Values (“Fundamentals”) . . . . .	13
2.2. Historical Correlation . . . . .	13
2.3. Industry Taxonomies . . . . .	14
2.4. Sell-side Equity Research Analysts . . . . .	17
2.5. News Article Co-occurrences . . . . .	20
<b>3. Pairwise Correlation Prediction</b> . . . . .	24

3.1. Set Representation and Interestingness Measures . . . . .	24
3.2. Data . . . . .	29
3.2.1. Stocks . . . . .	29
3.2.2. Analysts . . . . .	30
3.2.3. News Articles . . . . .	30
3.3. Predictiveness of Individual Datasets . . . . .	34
3.3.1. Selectivity Properties . . . . .	37
3.4. Method to Combine Datasets . . . . .	39
3.4.1. Controlling for Confidence . . . . .	41
3.4.2. Importance of Factors . . . . .	44
Predictors . . . . .	44
Confidence . . . . .	46
3.5. Summary . . . . .	47
<b>4. Hypergraph Partitioning . . . . .</b>	<b>48</b>
4.1. Motivation . . . . .	48
4.2. Background and Related Work . . . . .	51
4.2.1. Progression of Partitioning Algorithms . . . . .	51
4.2.2. Quality Measures & Balance Constraints . . . . .	52
4.2.3. Consensus Clustering . . . . .	54
4.3. Method . . . . .	55
4.3.1. Algorithm Outline . . . . .	55
4.3.2. Discount Cut . . . . .	55
4.3.3. Implementation & Complexity . . . . .	58
4.3.4. Drawbacks . . . . .	60
4.4. Experimental Results . . . . .	61
4.4.1. Performance with Imbalance . . . . .	62
4.4.2. Performance with Edge Size . . . . .	64
4.4.3. Real World Datasets . . . . .	64

4.4.4. Consensus Clustering . . . . .	66
4.5. Summary and Future Directions . . . . .	69
<b>5. Stock Groups . . . . .</b>	<b>70</b>
5.1. Evaluation Measures . . . . .	71
5.1.1. Average Pairwise Correlation . . . . .	71
5.1.2. Average Coefficient of Determination . . . . .	73
5.2. Methods . . . . .	74
5.2.1. Hypergraph Partitioning . . . . .	75
5.2.2. Conversion of Pairwise Data into Groups . . . . .	78
5.3. Results . . . . .	80
5.4. Summary . . . . .	84
<b>6. Applications . . . . .</b>	<b>92</b>
6.1. Diversification . . . . .	92
6.1.1. Experimental Setup . . . . .	93
6.1.2. Results . . . . .	95
Portfolio Selectivity . . . . .	95
Comparison with GICS . . . . .	97
Combination with GICS . . . . .	98
6.2. Long Position Hedging . . . . .	98
6.2.1. Experimental Setup . . . . .	100
6.2.2. Results . . . . .	102
6.3. Further Applications . . . . .	104
6.3.1. Comparative Analysis & Company Valuation . . . . .	105
6.3.2. Equity-Neutral Trading . . . . .	106
<b>7. Conclusion . . . . .</b>	<b>108</b>
7.1. Future Work . . . . .	109
7.1.1. Extraction of Relationships from News or Other Textual Data . . . .	109

7.1.2.	Measure Timeliness . . . . .	109
7.1.3.	Labeling Stock Groups . . . . .	110
7.1.4.	Hypergraph Partitioning Improvements . . . . .	111

# Chapter 1

## Introduction

Like many industries, finance has been continually transformed by new technologies. The New York Stock Exchange (NYSE) is often recognized as originating with the signing of the Buttonwood Agreement in 1792 (Geisst, 2012, Chapter 1). In these early days of trading, stock and commodity prices had to be delivered by person. Large numbers of people would be employed with the simple goal of communicating stock quotes as quickly as possible. Still, physical limitations meant information disseminated slowly. The introduction of the telegraph in the mid-1800s greatly reduced the time to send a message and meant trading could more readily be performed over larger distances. Soon after, the ticker tape machine was invented, eliminating the need for a trained receiver to transcribe signals. One of the best known machines is the Universal Stock Ticker, an early development by Thomas Edison that helped provide the funds to create his famous laboratory in Menlo Park.

In the twentieth century, technology and automation continued to be part of advances in the financial industry, sometimes by necessity. In 1968, paper stock certificates were still required to be physically delivered after a transaction. Increases in trading volume meant that back offices soon became overwhelmed, resulting in the “Paper Crisis” and NYSE actually stopped trading on Wednesdays for nearly the entire second half of 1968 so that back offices could recover. This led to the accelerated formation of the Central Certificate Service, and later the Depository Trust Company, to store and transfer stock ownership electronically. The Paper Crisis also served as impetus to form the world’s first fully electronic exchange, National Association of Securities Dealers Automated Quotations (NASDAQ), which replaced the previous over-the-counter (OTC) system of trading smaller stocks by phone (Wells, 2000). Older exchanges that previously used open outcry have been converting to electronic trading for years, including the London Stock Exchange (LSE),

which converted in 1986, and NYSE, which moved to a “hybrid” market in 2007. Meanwhile, electronic communication networks (ECNs), crossing networks and dark pools have become prevalent forms of Alternative Trading Systems (ATs).

The advent of new technologies have often brought controversy. “Black Monday” (Oct. 19th, 1987), the largest one-day decline in the Dow Jones Industrial Average (DJIA), is often blamed in part on “program trading”, which is a broad term but with regard to Black Monday usually refers to a trading method that had relatively recently been introduced and allowed execution of trades in a basket of securities contingent on certain conditions met. Traders often used program trading to implement “portfolio insurance” whereby a portfolio would be sold if stocks began to decline. The widespread use of portfolio insurance is believed to have caused a positive feedback system with sales by program trades triggering more and more sales by other program trades. (Carlson, 2007) More recently, use of high-frequency trading has been blamed for the “Flash Crash” of May 6, 2010, where the DJIA lost over 9%, then recovered within minutes (CFTC/SEC, 2010).

Still, technology and automation have been part of clear long-term beneficial trends. Jones (2002) finds that one-way commissions as a proportion of the trade’s value have decreased from approximately 0.25% in 1925 to 0.10% in 2000. Further, bid-ask spreads have decreased from approximately 0.65% in 1900 to 0.20% in 2000. Chordia et al. (2011) finds dramatic reductions continuing into the 21st century. These changes can be partially attributed to changes in regulation, especially the 1975 “May Day” deregulation of the brokerage industry by the Securities and Exchange Commission (SEC) and also decimalization in 2001. Yet, the influence of technology in the reduction of trading costs is beyond doubt. Even the controversial techniques of algorithmic and high-frequency trading show evidence of bringing greater trading volume and greater efficiency to the market (Hendershott et al., 2011).

This dissertation is intended to provide support to this continual progression towards automation and efficiency. More specifically, it seeks to consider the task of predicting company similarity, particularly as measured by stock price co-movements. Two output forms often used by financial practitioners are examined. The first is pairwise similarity where a scalar value is produced to quantify the similarity between two companies. This is
















Stock Sector   Holdings Detail >>		
	Portfolio (% of Stocks)	S&P 500 (%)
 <b>Cyclical</b>	<b>7.62</b>	<b>31.25</b>
 Basic Materials	0.00	3.21
 Consumer Cyclical	2.34	11.29
 Financial Services	5.18	14.79
 Real Estate	0.10	1.95
 <b>Sensitive</b>	<b>92.37</b>	<b>42.19</b>
 Communication Services	0.23	4.22
 Energy	0.00	10.89
 Industrials	4.50	10.00
 Technology	87.64	17.08
 <b>Defensive</b>	<b>0.01</b>	<b>26.56</b>
 Consumer Defensive	0.01	10.75
 Healthcare	0.00	12.37
 Utilities	0.00	3.44
 Not Classified	0.00	0.00

Figure 1.1: Example Morningstar Instant X-Ray. In this instance, the portfolio has a large exposure to technology.

analogous to the Pearson correlation of returns for two stocks (although this dissertation’s goal is prediction of similarity, not necessarily the quantification of historical similarity). The second form is a simpler representation where stocks are placed in clusters, with the goal that each stock is most similar to the other stocks in its cluster. These clusters often appear in practice as sector/industry classifications or taxonomies, with each company belonging to a group, like “Health Care”, “Energy”, “Consumer Staples”, etc.

The use of such similarity information in the financial community is pervasive. For instance, services like the Instant X-Ray from Morningstar<sup>1</sup> allow investors to view the weights of their portfolio across sectors with the goal of avoiding too much concentration in any single sector. (See Figure 1.1.) They might also be used to hedge stock positions in a equity-neutral fashion. For example, an investor might hold a stock position and fear that a sector or market decline will reduce the position’s value. Meanwhile, that investor might be unable to liquidate the position due to trading restrictions, such as black-out periods. To hedge the position, the investor could sell short a number of highly similar stocks, such as those in the position’s sector. So, if the market or that sector were to experience a general

<sup>1</sup><http://morningstar.com>

decline (or rise), any losses in the long position should be compensated by gains in the shorted stocks (or vice-versa). Underlying this strategy is a belief that the stocks will have highly similar reactions to market or sector events (e.g., a rise in the federal funds target rate or a decline in consumer sentiment). Thus, an accurate sector mapping is vital to the strategy.

Yet, creation of sector/industry taxonomies and other quantifications of company similarity has often been viewed as an expert task. Financial ratios and other metrics, such as the price-to-earning (P/E) ratio, return on assets (ROA), operating margin, etc., might be used to inform the expert, but the task of conceiving the groups, levels and definitions for the taxonomy are left to the expert.

This dissertation seeks to provide methods that compute company similarity and produce clusterings in a more automated fashion, while still offering quality that is “on par” or even exceeds expert-devised taxonomies. Further, this dissertation identifies two novel datasets that can aid in the computation of company similarity: equity analyst coverage and news article co-occurrences. The primary purpose of neither is to compute company similarity; yet, both convey rich information about company relatedness. This dissertation offers methods to convert these datasets into the more immediately usable forms: pairwise similarity and stock groups. The contributions of this dissertation are described more precisely in the next section.

## 1.1 Contributions

This dissertation’s contributions are grouped into three categories: new perspectives, methods and tools, and applications.

### 1.1.1 New Perspectives

**Sector/Industry taxonomies are not ground truth** Much research, particularly from the computer science community, takes the view that stock grouping is an expert task and follows a paradigm of devising a method, then offer a demonstration of agreement with sector/industry taxonomies to show validity of the method. (Examples



are: Doherty et al. (2005); Gavrilov et al. (2000)). This dissertation takes the view that improvements can be made beyond what is achieved by these taxonomies.

**“Don’t listen to what analysts say. Look at what they cover.”** Many studies (e.g., Welch (2000)) have found that analyst earnings estimates and stock recommendations subject to biases. Ignoring this primary function of analysts, this dissertation finds an alternative use that is powerful and surprisingly accurate: determining stock similarity. The strength of this feature stems from the fact that in order to facilitate efficiencies at the research firms employing the analysts, each analyst is typically assigned to a set of highly similar stocks.

### 1.1.2 Methods and Tools

**Entropy-constrained hypergraph partitioning** A method for constructing stock groups directly from the analyst or news data is described. The data are represented as a hypergraph and partitioning is applied such that agreement with the original dataset is maximized. In the case of analysts, the method seeks to achieve this maximization by minimizing the instances where an analyst’s covered stocks spans in multiple parts of the partition. In order to ensure balance roughly matches comparison groups, an entropy-constraint is imposed. The method can easily be extended to other domains using hypergraph representations, such as consensus clustering, where many researchers have been frustrated by the fact that most existing tools offer constraints that are focused on equal-sized partitions. In addition, a “discount-cut” heuristic is described that helps avoid a local optima problem that is frequently found in existing methods. The method is implemented in C++ and links to the source code are provided.

**Pairwise similarity computation through the cosine measure** Several interestingness measures from frequent itemset analysis are considered for use in computing a similarity value for pairs of stocks using the news and analyst datasets. The cosine measure is selected based on many desirable features relevant to the company similarity setting.

**Procedure for combining pairwise datasets** A method to find optimal weights among

datasets to use for prediction of future correlation is described. The method is used to combine historical correlation, analyst cosine values and news cosine values, with results that generally perform at least as good as the best performing single dataset and often even better. These performance improvements are particularly true for pairs of stocks with the highest similarity (i.e., the “top K”), which is important for a variety of applications, such as hedging and relative valuation. The most similar stocks are the best candidates for hedging and the most similar stocks would form the “peer group” for relative valuation.

**Pipeline approach to creating groups from pairwise similarity values** A method to form stock groups from pairwise similarity values is described that starts with a fast hierarchical clusterer, then applies improvements using a genetic algorithm. Links to source code for the methods are provided.

### 1.1.3 Applications

Numerous possibilities for using company similarity exist. Potential ideas are described in chapter 6 with the following two examined in depth.

**Long position hedging** Suppose an investor holds a position in a single stock and fears an impending sector or market decline will reduce the value of his/her position, but the investor is unable to sell the position or invest in derivatives with value derived from the price of the stock. Such situations are not uncommon. For example, executives are often highly restricted in trading in their company’s stock, even though much of their compensation (and therefore their wealth) may be in stock or stock options. Even non-executive employees with stock purchase plans may be restricted from trading during a vesting period. In such situations, hedging is vital to maintaining wealth. Using the pairwise similarity values computed in chapter 3, this dissertation demonstrates that by selling short a portfolio of most similar stocks, risk is reduced by using a combination of analyst, news and historical correlation data.

**Diversification** Whereas hedging reduces risk by seeking the most similar stocks and selling them short, diversification seeks to avoid similarity entirely. That is, a portfolio

is diversified if its constituents have minimum correlation and, therefore, reduced risk of simultaneous losses. This dissertation compares using pairwise similarity values to avoid similar stocks with standard approaches using sector/industry taxonomies that seek to achieve diversification by avoiding portfolios with concentrations in any single sector.

## 1.2 Dissertation Structure

The remainder of this dissertation is organized as follows. In chapter 2, prerequisite information is provided and datasets are described. Related work is also summarized. In chapter 3, methods are provided to quantify pairwise similarity using the datasets and the foundational task of correlation prediction is examined. In chapter 4, hypergraph partitioning is discussed and an algorithm is provided that performs well in partitioning imbalanced datasets. Subsequently, formation of stock groups is considered in chapter 5, both by using the datasets directly through the hypergraph representation and by using the pairwise values of chapter 3. Applications are examined in chapter 6 and chapter 7 concludes.

## Chapter 2

### Background & Related Work

Finance, particularly trading, is a peculiar area in that many resources are devoted to research, but a relatively small amount is published. Knowledge is often kept secret, since profiting from new ideas can more immediately be accomplished than with other fields. A breakthrough in medicine requires years of tests, government approvals and an apparatus to manufacture, distribute and market new medicines, whereas a trading insight may only require some start-up capital and perhaps software systems to achieve profitability. Still, there have been many advances driven by academics and enough of the industry is visible to infer some of its workings.

This chapter presents a literature review with the intention to provide some of the prerequisite knowledge required in later chapters and also to suggest this dissertation's place within the existing literature. The chapter begins by explaining and justifying correlation as a primary measure of company similarity. Next, historical correlation's predictive power of future correlation is discussed, followed by a primer on industry taxonomies, which serve as an expert-driven reference point by which this dissertation's methods can be compared. Finally, the two nontraditional datasets, analyst coverage and news article co-occurrences, are discussed.

#### 2.1 Correlation as a Measure of Similarity

The major goal of this dissertation is to quantify the “similarity” of companies. Depending on a given viewpoint, similarity could be measured in a variety of ways. A consumer might wish to know which companies offer similar products or services. For example, if the consumer needs to file taxes, s/he might wish to know the different companies that offer such services. From this perspective, similarity in the company “outputs” are important.

Conversely, a supplier, such as raw materials producer, might be interested in the “inputs,” specifically which companies might need aluminum, sulfur, rare earth metals, etc. In a separate scenario, a politician might categorize companies based on their tendencies to offer campaign donations. There may exist interrelationships in the results of these metrics. For example, many companies that are similar as measures by their “inputs” may also be similar in their “outputs”. Both train and car manufacturers will have similar inputs, while their outputs are also similar. Still, applying each of the different metrics will likely have different results, hence it is important to understand the intended audience for any similarity quantification and also how that audience will use it.

This dissertation focuses on financial practitioners and investors, specifically those who wish to use the similarity values for hedging, diversification or other investing applications (i.e., the applications to be described in chapter 6). To this audience, understanding how the stocks of certain companies will “co-move” is a vital interest. For example, if a clothing retailer has a decline in stock price, can it be expected that other clothing retailers will decline in the same time period? Further, can it be expected that clothing manufacturers will also decline? Such propensities for co-movements are often quantified by the Pearson product-moment correlation coefficient, which will henceforth be simply referred to as “correlation.” Suppose there are two stocks,  $x$  and  $y$ , with time series of stock prices

$$P_{x,0}, P_{x,1}, P_{x,2}, \dots, P_{x,T} \quad \text{and} \quad P_{y,0}, P_{y,1}, P_{y,2}, \dots, P_{y,T}$$

For simplicity, these prices are measured at even time steps and for most of this dissertation daily close prices will be used.<sup>1</sup> The one-period return at time  $t$  using stock  $x$  is

$$R_{x,t} = \frac{P_{x,t} - P_{x,t-1}}{P_{x,t-1}}$$

The mean return is

$$\bar{R}_x = \frac{1}{T} \sum_{t=1}^T R_{x,t}$$

and the standard deviation of returns is

$$\sigma_x = \sqrt{\frac{1}{T-1} \sum_{t=1}^T (R_{x,t} - \bar{R}_x)^2}$$

---

<sup>1</sup>Actual price changes are usually unsynchronized, which is not important for this dissertation, but is important in high-frequency scenarios.

Finally, the correlation is

$$\rho_{x,y} = \frac{1}{n-1} \sum_{t=1}^T \left( \frac{R_{x,t} - \bar{R}_x}{\sigma_x} \right) \left( \frac{R_{y,t} - \bar{R}_y}{\sigma_y} \right)$$

An intuitive reason that correlation is applied to returns rather than stock prices directly is that investors are typically interested in the gains or losses on assets rather than their absolute price levels. A second, more statistical reason is that residuals in the time series of prices will likely have a trend and be non-stationary, meaning time series parameters like mean and variance will change over time. Non-stationarity is also possible with returns, but effects tend to be much less dramatic and an assumption of a fixed mean and variance is much more reasonable than with prices. A related, intuitive argument is that if one views the time series of prices as a product of returns, then it is evident that early returns have much greater weight in the computation of correlation if prices are used:

$$\begin{aligned} & (P_{x,0}, P_{x,1}, P_{x,2}, \dots, P_{x,T}) \\ &= (P_{x,0}, P_{x,0} \cdot (1 + R_{x,1}), P_{x,0} \cdot (1 + R_{x,1}) \cdot (1 + R_{x,2}), \dots, \\ & \quad P_{x,0} \cdot (1 + R_{x,1}) \cdot \dots \cdot (1 + R_{x,T})) \end{aligned}$$

In the seminal work introducing Modern Portfolio Theory (MPT), Markowitz (1952) demonstrates correlation's vital importance in risk reduction. Markowitz argues that risk can be measured as the variance of returns deviating from expected return. By including more assets in a portfolio, risk can be reduced as long as the assets are not perfectly correlated (and assets are rarely perfectly correlated). To demonstrate this result, consider the variance ( $\hat{\sigma}$ ) of a portfolio of assets

$$\hat{\sigma}^2 = \sum_i w_i^2 \sigma_i^2 + \sum_i \sum_{j \neq i} \rho_{ij} w_i w_j \sigma_i \sigma_j$$

where  $i$  and  $j$  are assets in the portfolio and  $w_i$  is the weight of  $i$  in the portfolio, such that  $\sum_i w_i = 1$ . Since  $-1 \leq \rho_{ij} \leq 1$ , it must be the case that

$$\hat{\sigma} \leq \sum_i w_i \sigma_i$$

Hence, risk (as measured by the standard deviation of returns) is reduced by holding a portfolio of assets versus just holding a single asset as long as the portfolio of assets are

not perfectly correlated. This fundamental result provides a strong theoretical argument for the benefits of “diversification” - avoiding holding similar assets. Thus, prediction of future correlation is of great interest to investors, not only for its usefulness in achieving diversification, but for other reasons that will be described in chapter 6, such as hedging. Therefore, future correlation is a key metric that this dissertation will use to quantify performance.

### **2.1.1 Possible Alternatives**

#### **Cointegration**

There do exist other measures of similarity. In particular, cointegration tests are often used to measure if two or more time series are highly related. Cointegration means that if the individual time series are each integrated at a given order, a linear combination of the time series can be found with a lower order of integration. Murray (1994) provides a simple example of a drunk and her dog who have recently left a bar. The drunk wanders aimlessly, as in a random walk. If the dog is considered independently, it also appears to follow a random walk. However, examining the drunk and her dog together, they tend to follow the same trajectory. They sometimes separate, but there is a strong force returning each to the other. Similarly, the stock prices of two related companies may wander around the same path and a test for cointegration will capture this relatedness. (See Alexander (2001) for a more rigorous explanation of cointegration, illustrated with financial applications.).

A major difference between correlation and cointegration is that correlation measures how strongly the returns of two stocks move similar movements in the same time steps. With cointegration, the two stocks do not need to have “synchronized” movements, but rather must follow the same general trends - deviations are allowed but reversion to the same spreads are expected. So, the underlying desires must be considered when choosing between use of correlation or cointegration. One application that widely uses cointegration is pairs trading (Gatev et al., 2006; Elliott et al., 2005), wherein two stocks are first identified as cointegrated. So, whenever the prices of the two stocks drift apart, it can be expected that they will revert. Thus, a trading strategy can be implemented that sells short the stock that

has drifted higher in price and buys the stock that is lower in price. As the prices revert to their long-term spread, one can expect to profit regardless of general market conditions. If the market drops, gains in the short position will compensate for losses in the long position. Likewise, if the market rises, any losses in the short position will compensate for gains in the long position. Since the strategy relies on drifts and reversions, cointegration is an appropriate test.

At the same time, in a risk management setting, it may be important that movements in the assets occur in the same time period. Suppose an investor holds a long position in a stock index fund, but believes the market will suffer a decline. Rather than incur taxes by selling the index position, the investor decides to hedge using a short position in futures. The investor's bank agrees to fund the futures position using the stock index position as collateral. If the stock index and its futures do not remain tightly correlated, the investor may be forced to liquidate part of the index position. In this case, drifts are not desirable, so correlation is more likely an appropriate measure than cointegration.

This dissertation uses correlation as the predominate measure of similarity for several reasons.

- Correlation appears more often in traditional financial models, particularly the Capital Asset Pricing Model (CAPM). Since this dissertation focuses on the finance community, it is appropriate to use the measure most commonly used by that community.
- Correlation is generally easier to understand, whereas cointegration requires a much stronger mathematical background. Focusing on correlation should make this dissertation accessible to a wider audience.
- Correlation is an appropriate measure in many situations, such as the hedging scenario above. Many other applications exist, such as those to be discussed in chapter 6, so it is reasonable to use it.

Examination of cointegration and consideration of its applications are left for future work.



## **Financial Statement Values (“Fundamentals”)**

While this dissertation focuses on stock returns, other time series could be used to measure similarity between companies. In particular, values from financial statements, such as sales, earnings, research & development expense, etc. could be used to quantify the relatedness of companies. These are often called “fundamentals.” Prediction of these values is also important to many financial practitioners. For example, they can be important to a buyout firm that wishes to assess the value of a takeover target.

The drawbacks of financial statement values are that they are much less frequent than stock prices. Generally, financial statements are filed only quarterly with regulators. Different accounting methods also mean that the values have different meanings from company to company. In contrast, stock returns have much higher frequency and have the same meaning across companies. Moreover, these stock returns ultimately have greatest importance to investors because they represent the investor’s change in wealth.

## **2.2 Historical Correlation**

Since prediction of correlation is a main goal, a natural question is how predictive is simply using historical correlation? As a prelude, results in chapter 3 will show that it does have strong predictive power, although not necessarily more power than other datasets. For now, theoretical arguments both in favor and against its predictiveness will be considered.

Companies and the economy are constantly changing. Google began as a web-search company, but has expanded over time into other domains. It has developed a popular mobile operating system, Android, and even acquired Motorola’s wireless handset division in 2012. Thus, one might expect its relatedness to other telecommunications companies to have risen over time. At the same time, it has been a leader in the development of autonomous driving systems and one might expect its relatedness to automobile manufacturers to increase in the future. From this perspective, one can expect inter-company correlations to change as companies change their products and services. Still, these changes occur slowly and so the past should hold some predictive power for the future, particularly within the span of a few years.

The “stationarity” of stock correlations is an important question that has been considered in prior research since models, such as the Markowitz model described in section 2.1, often require ex-ante (i.e., future) correlations as input, but since these are unknown, ex-post (i.e., historical) correlations are frequently used instead. This approach implicitly assumes stationarity in the correlations. Contrary to these expectations, Cizeau et al. (2001) find that correlations increase in periods of high volatility – often when the diversification effects of low correlations are needed most. (Similar results are found in Reigner et al. (2011) and Preis et al. (2012).) Yang et al. (2006) finds similar results and also that international correlations have been increasing as the global economy becomes more integrated (a result also found by Cavaglia et al. (2000)). Tóth and Kertész (2006) has found correlations, on average, have increased over time. Their period of examination was 1993 to 2003, but this dissertation finds similar results through 2010 (see section 3.3). Each of these results indicate correlations are neither completely stationary, nor are they completely time-independent. Thus, one can expect historical correlation to have some predictive power, but not absolute predictive power.

## 2.3 Industry Taxonomies

In contrast to correlation, which is a purely numerical quantification of similarity, sector/industry taxonomies are generally constructed by experts. These experts create a hierarchy designed to partition the economy into groupings such that the groups are similar across some attributes or set of attributes. Examples of these attributes are production processes, products and services, and responses to economic factors.

The development of these taxonomies has a long history. The predominate classification system in the U.S. for much of the twentieth century was the Standard Industrial Classification (SIC), developed by the U.S. government in the 1930s. SIC Codes have 4 digits for each “industry.” These can be grouped by the first 3 digits or first 2 digits to indicate the “industry group” or “major group,” respectively. The major groups can also be mapped into 9 “divisions.” These levels of granularity allow the user flexibility when aggregating data from individual companies across industries. The SIC codes and hierarchy have been periodically updated to account for structural changes in the economy, with the last update

in 1987. In 1997, the North American Industry Classification System (NAICS) was jointly introduced by the Canadian, Mexican and U.S. governments with the intention of replacing SIC. NAICS has 6 digits, with grouping possible on digits 2 through 6, allowing 5 levels of granularity. Despite the introduction of NAICS, use of SIC codes is still pervasive, with the U.S. Securities and Exchange Commission (SEC) a notable user that has not transitioned to NAICS. SIC codes have also been used widely by academic studies due to their availability and long history.

Bhojraj et al. (2003) explain two major weaknesses of SIC codes for investment researchers which carry over to NAICS. First, the U.S. Federal Census Bureau establishes the taxonomy, but does not actually assign codes to companies. Company assignments are left to data vendors or whomever wishes to use the taxonomy. Guenther and Rosman (1994) show that even at the major group level (first two digits of code) there is 38% disagreement between the classification of SIC codes by two of the largest financial data providers, Compustat and the Center for Research in Security Prices (CRSP). Their work also shows that Compustat’s intra-group price correlation is significantly larger than CRSP. Differences such as these can add ambiguity to any research results. The second weakness for financial practitioners is that these taxonomies have a “production-oriented, or supply-based conceptual framework” (Katzen, 1995), which were designed to support the needs of government statistical agencies wishing to report on the economy (Arbuckle, 1998). Private investing was less a concern in their design.

With some of these issues in mind, Fama and French (FF) present a re-mapping of SIC codes into 48 industry groups in their study of industry costs of capital Fama and French (1997). French provides additional mappings to 5, 10, 12, 17, 30, 38 and 49 groups through his data library French (2012). The FF scheme is intended to form groups that are likely to share risk characteristics and often appears in academic studies (Bhojraj et al., 2003).

Targeting financial practitioners, S&P and MSCI devised the Global Industry Classification Standard (GICS), which was announced in 1999 and replaced S&P’s previous industry classification methodology in 2001 (Maitland and Blitzler, 2002). The methodology classifies companies “based primarily on revenues; however, earnings and market perception are also considered important criteria for analysis” (MSCI / Standard & Poor’s, 2002). As seen in

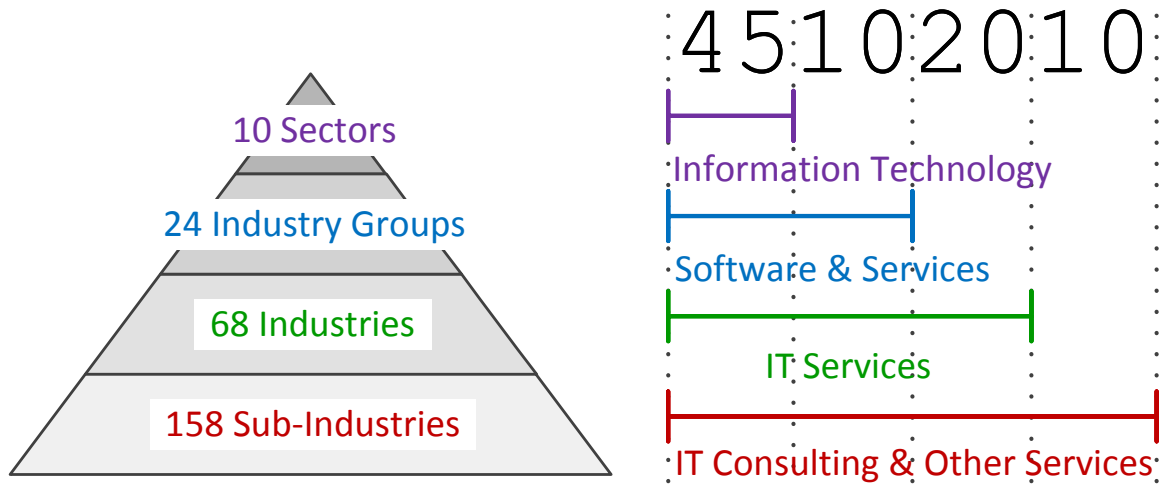


Figure 2.1: GICS Structure (as of Jan. 1, 2010). As an example, the code for International Business Machines (IBM) is shown on the right along with the textual names corresponding to its classification.

Figure 2.1, a single hierarchy is devised based on an 8-digit encoding that can be used to group companies at the sector, industry group, industry and sub-industry level using the first 2, 4, 6 and 8 digits of the code, respectively. This hierarchy is applied to companies globally in order to simplify cross-border comparisons.

Bhojraj et al. (2003) compare SIC, NAICS, FF and GICS by computing, for each stock group in each taxonomy, an average time series for stock-price returns and for seven financial ratios, such as price-to-earnings or return-on-equity. Regression is performed for each stock against its group's average and the taxonomies are compared by their average  $R^2$  values. They find higher  $R^2$  for GICS in nearly every aspect of the comparison and conclude that GICS is a superior classification system.

Chan et al. (2007) provide an alternate methodology that computes the difference between the pairwise intra-group and inter-group correlation, with a higher difference attributable to better groupings. (See section 5.1.1.) They suggest this approach should be preferred because it is more applicable to portfolio analysis and risk management. Their work compares FF, GICS and a hierarchical clustering algorithm that groups stocks on the basis of a five-year history of returns correlations. They find hierarchical clustering

performs better in the training period, but underperforms when used in subsequent periods. Their results show GICS achieves the highest difference in correlation with the fewest number of groups, thus outperforming FF and historical clustering. Using a similar methodology, Vermorken (2011) compares GICS with its main commercial competitor, the Industry Classification Benchmark (ICB), which is the product of Dow Jones and FTSE. Though differences exist, Vermorken ultimately finds the taxonomies largely similar for his sample of large-cap stocks. Due to these prior works demonstrating the strength of GICS, it is used as the predominate sector/industry taxonomy for purposes of comparison in this dissertation.

It must be noted that historical correlation has been used in previous research to identify structure in stock markets. For example, Mantegna (1999) uses a correlation matrix to form an economic taxonomy. In another work, Micciche' et al. (2005) perform hierarchical clustering on stocks in the NYSE and suggest results have agreement with the Standard Industrial Classification (SIC). In chapter 5, this dissertation goes further by forming groups that have higher quality than SIC on two different measures. Further, correlation is combined with other datasets to find even greater performance in forming stock groups.

Finally, to emphasize a point made in section 1.1.1, this dissertation will take a fundamentally different view than much existing research: improvements can be made beyond the expert approaches, rather than seeking agreement with them as verification. GICS's wide spread use in the financial community and continued adoption over the existing SIC system effectively make it a "state-of-the-art" system, so any results that match or exceed the quality of GICS should be meaningful to the financial community.

## 2.4 Sell-side Equity Research Analysts

A major dataset used in this dissertation is "analysts." Outside the context of this dissertation, an "analyst" can be someone performing a variety of functions, but in this thesis, they are a specific set of individuals: sell-side equity research analysts. We more precisely define these individuals and their functions below.

Financial firms are often split into the "buy-side" and the "sell-side". The buy-side represents mutual funds, pension funds, hedge funds, endowments and other institutions

with money to invest. The sell-side consists of brokerages, investment banks, trade-clearing corporations and other institutions that offer services to the buy-side (Hooke, 2010, Ch. 2). One of these services is research, wherein firms will offer reports or verbal consultation on a variety of topics from general economic outlook to evaluations of individual assets. A large portion of the research generated by these firms tracks individual companies, estimating their earnings and offering investment recommendations for those individual companies. In order to perform such research, each firm offering this service will typically employ a large number of analysts, often called “sell-side equity research analysts”.

The research firm will typically assign each of these analyst to “cover” a set of companies such that the companies in the set are highly related. For example, an analyst may be assigned to cover restaurants, or clothing retailers (Valentine, 2011, Ch. 2). Some reasons for the firm to ensure the companies covered by an analyst are highly related are

- Each analyst can specialize in a particular area, rather than dividing efforts among many areas. For example, if an analyst had to cover Microsoft and Google, the analyst must have knowledge of the technology industry, but if the companies were Microsoft and Exxon, the analyst must have knowledge of both technology and energy industries. Further, these knowledge requirements are not static. As the industries evolve, so must the analyst’s knowledge. Hence, covering multiple areas is much more time-consuming.
- In order to effectively cover a particular company, the analyst will likely have to understand its competitors. It makes sense to have that same analyst also cover those competitor companies, rather than paying another analyst to cover them. These competitor companies will likely be the ones most similar to the original company.
- A research firm is a business. Having more efficient analysts will mean lower costs for a firm, giving it a competitive advantage over other firms.

When a firm makes such coverage assignments, it performs a partition of the universe of stocks. For the reasons above, it can be expected that those stocks will be highly similar, and this dissertation seeks to utilize this data as will be seen in the subsequent chapters.

Investors do listen to these analysts and their opinions can greatly influence the price of a company's stock (U.S. Securities and Exchange Commission, 2010). This importance has meant that analyst recommendations and earning estimates have been well-tracked. For example, the Institutional Brokers' Estimate System (I/B/E/S), our primary analyst data source, has tracked analysts since 1976 (Thomson Reuters, 2014). Their apparent importance to the investment community and the availability of this data has lead to a plethora of academic research on sell-side analysts. Much of this research focuses on their accuracy and the identification of biases. Some research suggests they do have investment value. For example, Barber et al. (2010) find that a portfolio following the literal levels of ratings (e.g. buy, hold, sell) generates positive abnormal returns. They find the same of a portfolio tracking revisions, which is one that buys stocks that are upgraded and shorts stocks that are downgraded. This occurs even on upgrades (downgrades) that maintain a negative (positive) level, such as an upgrade (downgrade) from strong sell (strong buy) to sell (buy). They find the greatest positive abnormal return when the rating and revision provide a consistent indicator and have the greatest strength. That is, a portfolio constructed of upgrades from hold to strong buy would outperform a portfolio constructed of upgrades from buy to strong buy, or of upgrades from strong sell to hold. These findings indicate that analyst stock recommendations do have predictive value.

At the same time, their accuracy has long been disputed. Nearly a century ago, Cowles, III (1933) found that contemporary stock market forecasters did worse on average than pulling randomly from decks of cards to choose predictions and dates for each prediction. More recently, "herding" has been observed, whereby analysts tend to follow each other's ratings more often than could be explained by chance (Welch, 2000). Moreover, in the early 2000s, several major research firms were accused of providing undeserved positive ratings to companies that paid for investment banking services from that firm. These apparent conflicts of interest resulted in the Global Analyst Research Settlement with several large firms forced to pay fines and erect physical barriers between their investment banking and research departments (Barber et al., 2007). Even research performed by this dissertation's author (along with a co-author) indicates the accuracy of analysts is very limited (Yaros and Imieliński, 2013c).

One major advantage of this dissertation’s use of the analyst data is that it bypasses this contention surrounding analyst accuracy. Since we use only the analysts’ coverage sets, the correctness of earnings estimates and stock ratings are not important. Instead, this dissertation mainly relies on the assumption that firms will assign analysts to similar companies, which seems reasonable since it can be assumed that the firm is seeking to minimize costs.

Interestingly, this perspective on the data has previously had little attention. In past literature, we find Ramnath (2002) to be the earliest use of analyst coverage to determine stock similarity. The author focused on how a company’s earnings announcement affects forecasts for other companies. The study wished to group stocks by industry, but recognized that Clarke (1989) and Guenther and Rosman (1994) had found issues with the industry classification scheme that was predominate at the time, the Standard Industry Classification (SIC). Ramnath then used a heuristic method to form groups where every stock in the group was covered by at least five analysts covering every other stock in the group. While Ramnath recognized that analyst coverage can be useful to determine groups, it was not the focus of his study. His heuristic method is somewhat arbitrary and, because of the five analyst threshold and other reasons, many stocks were completely omitted. This dissertation seeks to examine the use of this data in much greater depth and using more robust methods.

## 2.5 News Article Co-occurrences

While interest in equity research has led to a market that can support a large number of research firms and sell-side analysts, an even larger dataset lies in news articles. When a writer creates a news article that contains a company or set of companies, that writer generally communicates rich information about the companies, albeit only a sliver of the total information. In particular, they often convey much information about the relatedness of companies. Consider the following snippets from New York Times articles involving Wal-Mart (an American discount retailer with many stores in various countries):

1. *The [advertising] agency changes were part of a strategy shift at Wal-Mart, the nation’s largest retailer, to compete more effectively against rivals like Kohl’s, J. C. Penney*



*and Target (Elliott, 2006).*

2. *Procter & Gamble, the consumer products company, reached an agreement yesterday to acquire the Gillette Company, the shaving-products and battery maker, ... The move is a bid by two venerable consumer-products giants to strengthen their bargaining position with the likes of Wal-Mart and Aldi in Europe, which can now squeeze even the largest suppliers for lower prices (Sorkin and Lohr, 2005).*
3. *BUSINESS DIGEST ... Federal regulators wrapped up the first set of public hearings on Wal-Mart's request to open a bank, but gave scant indication of how they might rule on the company's application. ... Stocks tumbled as strength in the commodities market fed inflation fears and stifled investors' enthusiasm over upbeat first-quarter earnings from Alcoa (NYT (2006)).*

In the first snippet, several companies are identified as competitors to Wal-Mart. In a diversified portfolio, it would make sense to avoid large positions in several of these stocks because the companies face similar risks. For instance, a drop in consumer spending would likely affect all retailers.

In the second snippet, we see that Procter & Gamble and Wal-Mart hold different locations in the same supply chain. While the article clearly mentions a battle between the companies to extract more value in the supply chain, the profitability of each company is again linked to similar risks, such as a drop in consumer spending.

In the third snippet, we see Wal-Mart is mentioned together with Alcoa (a producer of aluminum), but there is no real relation between the companies presented in the article, other than the fact they had notable events occurring on the same day and, therefore, appear together in a business digest.

This dissertation hypothesizes that some signal of company relatedness can be captured by simply examining the co-occurrences of companies in news articles, despite presence of “noise,” such as in the case of the third snippet above. After controlling for the fact that some companies simply appear in news more than others, it can be expected that more co-occurrences should mean greater similarity for any set of companies. Additionally, the abundance of news articles and their widespread availability, particularly due to a trend

towards publication on the Internet instead of news print, means there exists a large and accessible dataset able to overcome the effects of noise.

Broadly examining existing research, the study of news and finance have intersected on a number of topics, including the speed of investor reactions to news stories (Klibanoff et al., 1998) and the effects of media coverage (or lack thereof) on stock prices (Chan, 2003; Fang and Peress, 2009). Another area is sentiment analysis, which has been applied to measuring the impact of pessimism on stock price and volumes (Tetlock, 2007). Sentiment analysis and use of other textual features have further been applied to create numerous signals for trading strategies (Li and Wu, 2010; Zhang and Skiena, 2010; Hagenau et al., 2012). Yet, those threads of research tend to focus on the use of news to predict the movements of single stocks, sectors or markets, rather than the relatedness and consequential co-movements of stocks.

The smaller thread of research more similar to this dissertation is the use of news and textual data to extract inter-company relationships. In a seminal work, Bernstein et al. (2002) use ClearForest software (a pre-cursor to the Calais service used in this article) to extract entities from a corpus of business news articles, which are then cleaned for deviations in company naming (i.e., I.B.M. vs IBM). Bernstein et al. then visualize the data with a network structure where edges are drawn between company vertices wherever the count of co-occurrences exceeds a set threshold. They highlight clusters in the network that appear to match common perceptions of industries, then develop a notion of “centrality” to measure a company’s importance to an industry. They further develop a cosine measure for the inter-relatedness of two pre-designated industries based on relatedness of their respective companies, as determined by news co-occurrences. As Bernstein et al. concede, the work is limited by its small set of articles covering only four months in 1999, where the dot-com bubble led to significantly high numbers of articles containing technology companies. Further, the results rely on the reader to judge whether the industries and relatedness measurements are reasonable, rather than offering verification through external measures of company similarity, such as stock-return correlation.

Other researchers have also considered use of news or other textual information to determine various aspects of relatedness between companies. Ma et al. (2009) construct a

network derived from news articles appearing on Yahoo! Finance over an eight month period. If the same article appears under the news pages for two different companies, a link is constructed between the two companies in the network. Multiple articles increase the weight of each link. Ma et al. then use in-degree and out-degree measures as features for binary classifiers that seek to predict which company in a pair has higher revenue. Jin et al. (2012) similarly construct networks based on co-occurrences in New York Times articles, but instead study the evolution of networks over time and use network features along with regression models to predict future company profitability and value. Rönqvist and Sarlin (2013) suggest bank interdependencies can be inferred from textual co-occurrences, rather than the two traditional data sources, co-movements in market data (e.g., CDS spreads), which are not always efficient, and interbank asset and liability exposures, which are generally not publicly disclosed. They exemplify their approach using a Finnish dataset and examine the temporal changes in the network, particularly following the Global Financial Crisis. Bao et al. (2008) present a method for extracting competitors and competitive domains (e.g., laptops for Dell and HP) that essentially uses a search engine to gather articles and then uses some sentence patterns to identify competitors. Hu et al. (2009) describe a system that extracts companies and corresponding relationships from news articles using predefined rules. They suggest an approach to detect events, such as acquisitions, by observing changes in the strength of relationships over time.

This dissertation differs from previous research by focusing on more robust evaluation of the usefulness of news, especially by examining its predictiveness of future correlation. This dissertation also evaluates its usefulness in hedging and diversification, which are of direct interest to financial professionals.

## Chapter 3

### Pairwise Correlation Prediction

As stated in section 2.1, a major goal of this dissertation is to identify data and develop methods to improve correlation prediction. In particular, we focus on the most similar companies (i.e., most correlated companies) since these are the most important in many applications, such as hedging and relative valuation. This chapter focuses on this task of similarity prediction, as measured by correlation. First, in section 3.1, cosine is identified as an appropriate measure to determine the similarity of two companies using the news and analyst datasets. Next, section 3.2 describes many of the datasets used to perform experiments in this chapter and throughout the remainder of this dissertation. In section 3.3, we measure the predictiveness and other properties of the analyst, correlation and news datasets. Finally, section 3.4 consider approaches to combining the datasets and ultimately find performance that is at least as good as the best individual dataset and often better, particularly for the pairs of most similar companies, particularly for the most similar pairs of stocks (i.e., the “top K”).

#### 3.1 Set Representation and Interestingness Measures

A first step towards computing similarity from the analyst and news datasets is to determine a representation of the data. Several possibilities exists. For example, a bipartite graph could be constructed with analysts or news stories forming the nodes as one part of the partition and companies forming the nodes of the other part. Another alternative would be to create a logical matrix (i.e., Boolean matrix) with companies forming the rows and analysts or news articles forming the columns. The cells of the matrix would be 1 wherever its corresponding company was present in the given article or covered by the given analyst.

In this chapter, a set representation is used. Using the analyst dataset, for each company,

Table 3.1: Hypothetical Analyst Coverage

Company	Symbol	Analysts Covering Company
Chipotle	CMG	Alice, Bob, Carol
Darden Restaurants	DRI	Bob, Carol, Dan
McDonald's	MCD	Alice, Bob
Netflix	NFLX	Frank, Mallory
Panera	PNRA	Alice
Yum! Brands	YUM	Bob, Carol, Dan, Oscar, Peggy

Table 3.2: Jaccard Similarity Values

	CMG	DRI	MCD	NFLX	PNRA	YUM
CMG	-	0.500	0.667	0.000	0.333	0.333
DRI	0.500	-	0.250	0.000	0.000	0.400
MCD	0.667	0.250	-	0.000	0.500	0.167
NFLX	0.000	0.000	0.000	-	0.000	0.000
PNRA	0.333	0.000	0.500	0.000	-	0.000
YUM	0.333	0.400	0.167	0.000	0.000	-

a set is formed of the analysts that cover the company. With news articles, for each company, a set is formed of the news articles that the company appears in. Thus, to quantify the similarity of two companies, one must quantify the similarity of their sets.

Consider Table 3.1. A simple measure of similarity for two companies might be to find the size of their overlap - i.e., the count of analysts covering both companies. For example, Chipotle and Darden would have overlap of two, while Chipotle and Netflix have overlap zero, indicating no similarity. However, overlap seems incorrect when considering Chipotle and Darden have the same value of two as Chipotle and Yum! Brands. Intuition suggests Chipotle appears should have less similarity with Yum! Brands because, in comparison to Darden, a smaller proportion of analysts covering Yum! Brands are also covering Chipotle. Hence, it is important to “normalize” by the sizes of the sets. At the same time, the

Table 3.3: Cosine Similarity Values

	CMG	DRI	MCD	NFLX	PNRA	YUM
CMG	-	0.667	0.816	0.000	0.577	0.516
DRI	0.667	-	0.408	0.000	0.000	0.775
MCD	0.816	0.408	-	0.000	0.707	0.316
NFLX	0.000	0.000	0.000	-	0.000	0.000
PNRA	0.577	0.000	0.707	0.000	-	0.000
YUM	0.516	0.775	0.316	0.000	0.000	-

penalty for having different levels of coverage should not be too severe. Both Panera and McDonald’s have analysts that are subsets of the analysts of Chiptole. This may reflect lower levels of coverage for the companies, rather than dissimilarity.

The quantification of set similarity is a well-studied area, particularly in frequent itemset and association rule mining where such measures are often called “interestingness measures.” Tan et al. (2005) provide a strong review of many of the measures used in literature and also identify several important properties common to some measures. Two properties important to this dissertation’s study are

**Symmetry** Company A’s similarity with company B should equal company B’s similarity with company A. This matches the objective, correlation, which will compute to the same value regardless of the order that the time series are considered. Many measures do not have the symmetry property, including conviction, mutual information, etc.

**Null Addition** In the example from Table 3.1, suppose more analysts exist than are present in table. If these analysts do not cover any of the stocks shown, then they should have no effect on the similarity values. That is, the measures should not be affected by “null addition.”

Only two symmetric measures out of nine considered by Tan et al. (2005) have the null addition property: the Jaccard index and cosine.

Let  $W_i$  represent the set of analysts covering a particular stock  $i$  and  $W_j$  represent the set of analysts covering a different stock  $j$ . The Jaccard index is

$$\mathcal{J}_{ij} = \frac{|W_i \cap W_j|}{|W_i \cup W_j|} \quad (3.1)$$

The measure was originally published by Jaccard (1901), but is sometimes called Tanimoto similarity due to its later publication by Tanimoto (1957).

The cosine is

$$\mathcal{C}_{ij} = \frac{|W_i \cap W_j|}{\sqrt{|W_i| \cdot |W_j|}}$$

This measure sometimes appears in literature as the Ochiai coefficient due to Ochiai (1957).

Following our initial thoughts above about a good similarity measure, both Jaccard and cosine measures contain the set overlap (i.e., intersection) in the numerator, but have different ways of normalizing (i.e., different denominators). Recalling the concern about different levels of coverage, cosine is a better choice because the effects of the set sizes in the denominator are dampened. In fact, cosine receives its name from its relationship to its geometric counterpart. The cosine between two vectors measures the angle between them, regardless of their magnitudes. Likewise the cosine interestingness measure quantifies the similarity between two sets, controlling for set sizes. Tables 3.2 and 3.3 provide the computed Jaccard and cosine similarity values, respectively, for the analyst coverage shown in Table 3.1. Note that Panera is more similar to Chipotle relative to McDonald's with the cosine measure than with the Jaccard measure, matching the desire that Panera not be penalized too severely for less coverage.

To further illustrate the need to control for variance in set sizes, consider Figure 3.1, which displays cross-sectional boxplots of the number of analysts covering each company. In the figure, the circle-dot represents the median number of analysts covering a company for the given year. The top and bottom of each thick bar represent the 75th and 25th percentiles, respectively. The top and bottom of each thin line represent the maximum and minimum, respectively. Correspondingly, Figure 3.2 displays the the number of times each company is mentioned in articles, across all article datasets (to be described in section 3.2.3). As can be seen in both Figures 3.1 & 3.2, the number of analysts and news mentions varies greatly (note the logarithmic scale) with imbalances more severe for news articles. This may

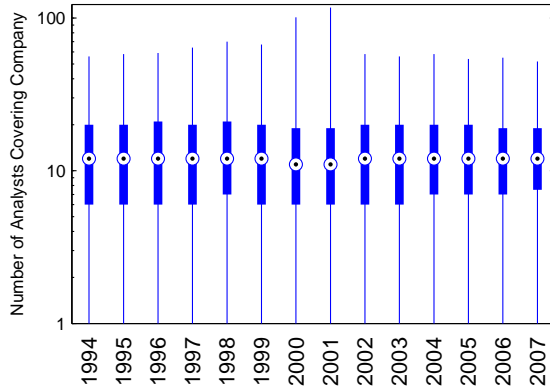


Figure 3.1: Analyst Coverage Counts

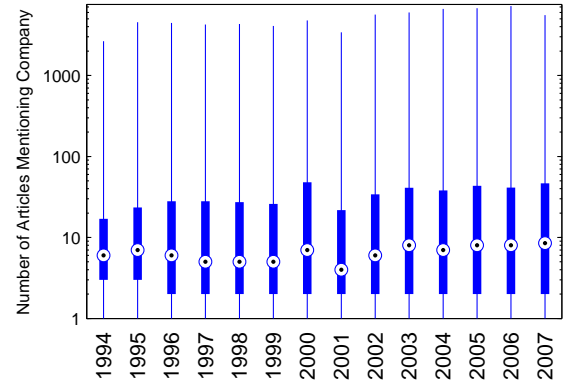


Figure 3.2: News Article Mention Counts

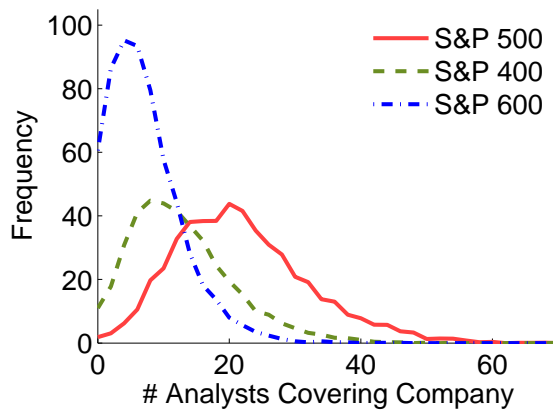


Figure 3.3: Analyst Coverage By Company Size

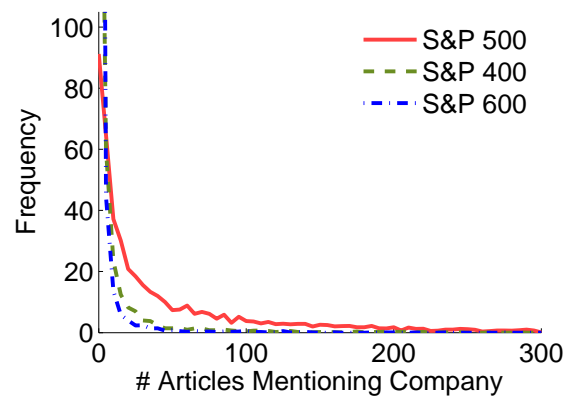


Figure 3.4: News Mentions By Company Size

occur because some companies are much larger and simply have more analyst coverage and news mentions due to their size. Figure 3.3 displays the number of analysts covering each stock averaged over years 1995 to 2008 for S&P 500, 400 and 600 stock, while Figure 3.4 displays the number of mentions in news. As will be discussed in section 3.2.1, the S&P 500 consists of five hundred of the largest capitalization stocks in the U.S. The S&P 400 consists of medium capitalization (i.e., “medium-cap”) stocks, while the S&P 600 consists of small-cap stocks. From the figures, it is evident larger companies tend to get more attention. However, company size does not account for all variance. Some small-cap stocks evidently get more analyst coverage and more mentions in news than some large-cap stocks. Intuitively, the variances in news article mentions can occur for a variety of reason. For example, some companies may have generated much controversy (e.g., Enron was frequently



in the news at the end of 2001 due to accounting fraud, but was previously less mentioned.) Analyst coverage can also vary for a variety of reasons, particularly demand from brokerage clients and potential for large price changes (e.g., mergers expected to occur). Regardless of the reasons, it is clearly important to control for the variations in set sizes - a need addressed by the cosine similarity measure.

In this section, cosine has been justified as a measure of similarity between two companies using analyst and news data. Shortly, its predictive power of future correlation will be examined, but first an overview of the actual datasets and the processing required to compute the cosine will be provided in the next section.

## **3.2 Data**

### **3.2.1 Stocks**

For stock returns, a dataset from the Center for Research in Security Prices (CRSP) is used which offers a daily return computation that includes splits, cash and stock dividends, and other distributions. The CRSP dataset also provides other corporate actions like delistings, mergers and spin-offs, so it is possible to have a complete picture of the active companies at any point in time. This property is critical to avoiding survivorship bias (Elton et al., 1996), which is an effect where results from many previous studies have been shown to have upward bias because they only examine companies that are active at the time of the study. Companies that went bankrupt or have otherwise been delisted are often excluded. This dissertation makes efforts to avoid such biases by including such stocks.

For the universe of stocks, the broad market S&P 1500 index is used, which is composed of the S&P 500 large-cap, S&P 400 mid-cap and S&P 600 small-cap stocks. Compustat, a product of S&P, is used to determine the composition of the S&P 1500 each year from 1996 to 2010. In order to avoid the aforementioned survivorship bias, the index compositions are fixed at the beginning of each year. Wherever possible, delisted stocks are included in computations by weighting them by the number of days they were active during the year. In chapter 5, Compustat will be further used to identify the GICS sector/industries for each company. Compustat will also be used for SIC codes, except wherever they are missing, in

which case CRSP will be used.<sup>1</sup>

### 3.2.2 Analysts

The data used to track analysts and their coverage comes from the Institutional Brokers Estimate System (I/B/E/S), which is currently owned by Thomson Reuters. Since 1976, I/B/E/S has recorded earnings estimates and recommendations from major brokerages.<sup>2</sup> The dataset differs from many others (such as FirstCall - acquired earlier by Thomson) by recording the analyst who made the recommendation or earnings estimate, not just the firm. Analysts are tracked throughout their career by a unique identifier which remains constant regardless of name changes. Actual coverage assignments are not recorded, so this dissertation instead slices the data into individual years and considers an analyst to be covering a stock if the analyst made at least one earnings estimate for that stock in the past year. Analysts estimate earnings for each financial quarter and their estimates are frequently updated as the company's earnings announcement approaches, so the using the estimates provides a strong proxy for the actual coverage assignments.

### 3.2.3 News Articles

The collection of news articles is taken from two corpora at the Linguistic Data Consortium (LDC)<sup>3</sup>. The first is the New York Times Annotated Corpus<sup>4</sup>, which contains over 1.8 million articles published by the New York Times (NYT) from January 1, 1987 to June 19, 2007. The second is English Gigaword Fourth Edition<sup>5</sup>, which contains articles from the following five newswire services<sup>6</sup>:

---

<sup>1</sup>This same prioritization of data sources for SIC codes is used elsewhere, such as Bhojraj et al. (2003).

<sup>2</sup>Ljungqvist et al. (2009) found evidence that I/B/E/S data had changed over time in ways that would improve the recommendation accuracy of some analysts, suggesting some inappropriate data modifications may have occurred. This dissertation uses data obtained after these modifications should have been corrected or reverted by I/B/E/S. Additionally, such modifications should not affect this study because earnings estimates are used. Further, their estimate values (i.e., their accuracies) are not important our study.

<sup>3</sup><https://www ldc.upenn.edu>

<sup>4</sup><http://catalog ldc.upenn.edu/LDC2008T19>

<sup>5</sup><http://catalog ldc.upenn.edu/LDC2009T13>

<sup>6</sup>The English Gigaword Fourth Edition also contains newswire articles from NYT, which we exclude since the separate NYT corpus is already being used.

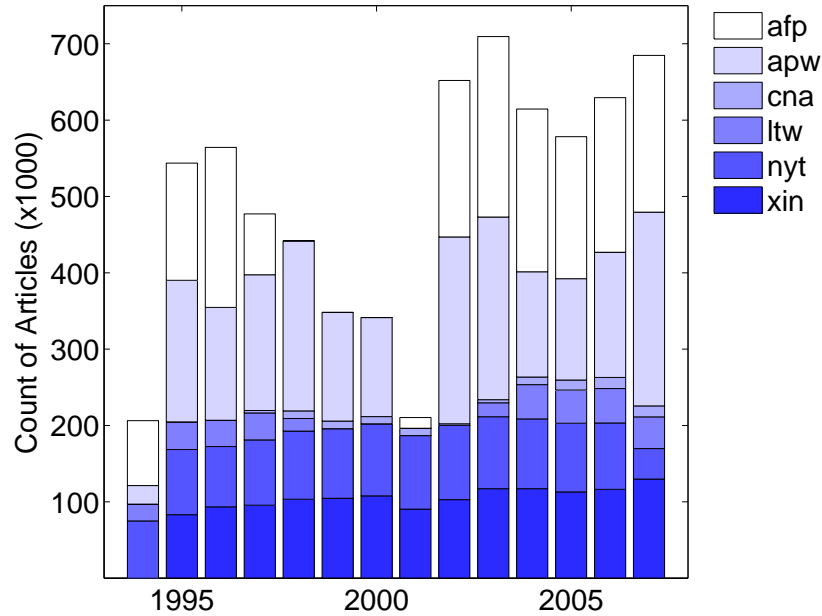


Figure 3.5: Articles Per Year

**AFP** Agence France-Presse

**APW** Associated Press Worldstream

**CNA** Central News Agency of Taiwan

**LTW** Los Angeles Times / Washington Post

**XIN** Xinhua News Agency

The data collection at LDC for most of the newswires was performed via dedicated lines that recorded article text and meta-data in real-time, although some articles were later received (or recovered) in bulk. Due to various collection issues, there are large gaps in the collection for particular newswires. There are also periods where fewer articles were collected due to changes in collection methods. Figure 3.5 depicts the number of articles from each source, per year<sup>7</sup>.

To extract company names from articles, we use Calais, which is developed and maintained by ClearForest<sup>8</sup>, a group within Thomson Reuters. The free OpenCalais<sup>9</sup> web service allows users to submit text and receive back annotations. Company name detection is a main feature, and is heavily used in this dissertation.

<sup>7</sup>A small number of articles contained non-English characters and could not be processed by OpenCalais. Figure 3.5 depicts only processed articles.

<sup>8</sup><http://www.clearforest.com>

<sup>9</sup><http://www.opencalais.com>

Table 3.4: OpenCalais Performance on 100 Sample Articles

No. Companies in Text	288		
True Positives	213	$F_1$ score	0.796
False Positives	34	Precision	0.862
False Negatives	75	Recall	0.740

To quantify OpenCalais error rates, we randomly selected 100 NYT articles and manually marked companies in the text. We then computed precision and recall as shown in Table 3.4. Calais does reasonably well at the difficult task of identifying companies, including differentiating those companies from non-profit organizations (e.g., Red Cross) or governmental agencies (e.g., Air Force). Some examples of false positives are shown in Fig. 3.6, where the words alleged to be companies by OpenCalais are outlined in boxes. In example (1), a region was misidentified as a company. Examples (2) & (3) demonstrate a common problem, wherein only part of the company name is identified as a company, or other words are combined into the company name (possibly from other companies). We did not attempt to perform any correction, again with the expectation that such errors will amount to noise and enough articles will overcome any problems. The worst case occurs when a false positive has the name of an actual company. For example, if the fruit “apple” was somehow identified by OpenCalais as the technology company, Apple. We never observed such situations, although it is possible they did occur. For most false positives, the misidentified text would not be included in our analysis because it simply would not link to our universe of stocks.

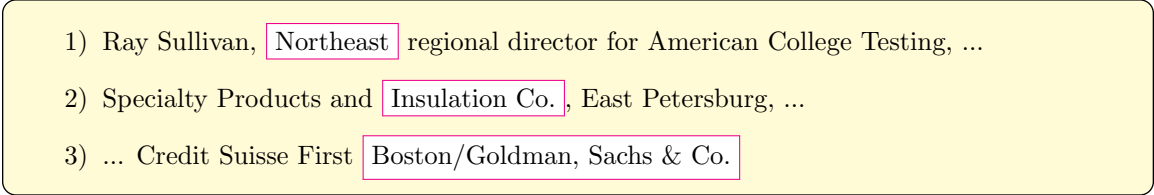
- 
- 1) Ray Sullivan, Northeast regional director for American College Testing, ...
  - 2) Specialty Products and Insulation Co., East Petersburg, ...
  - 3) ... Credit Suisse First Boston/Goldman, Sachs & Co.

Figure 3.6: Examples of False Positives

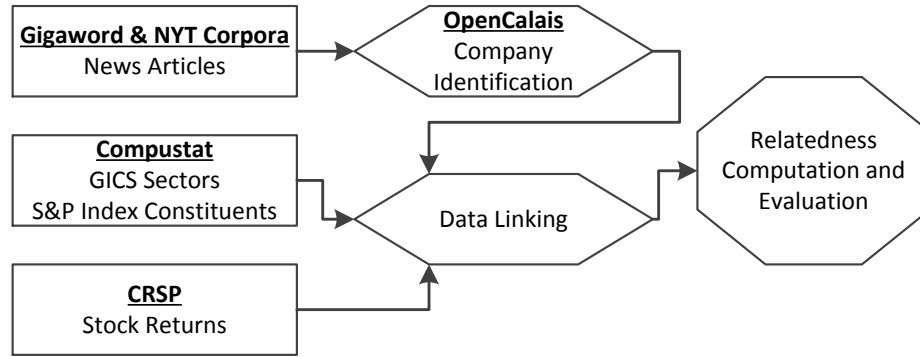


Figure 3.7: News Data Processing Overview

A significant task was linking the news article companies because the same company may be referenced in multiple ways. For instance, “DuPont,” “EI DuPont,” “E.I. Du Pont De Nemours” and “E.I. du Pont de Nemours and Company” are all aliases for the same company. Calais does offer company “resolutions,” where these multiple aliases are resolved to a single company name and Reuters Instrument Code (RIC). However, these resolutions do not account for the time period of the article. For example, “Mobil” will resolve to ExxonMobil, which is not helpful if the article is prior to the 1998 merger of Exxon and Mobil. Therefore, we use only the original company names identified by Calais, not their resolutions.

To link the identified companies to the other datasets, we use a manually constructed mapping of company aliases to CRSP permno (CRSP’s unique identifier). Each entry of the mapping includes beginning and ending dates of validity to avoid mapping to incorrect companies for the given article’s date, such as in the ExxonMobil example above. We further use a series of standardization rules on the raw company names, such as removal of periods, consolidation of abbreviation characters (i.e. “A. T. & T.” → “AT&T”), removal of suffixes (i.e., remove “Co,” “Company,” “Inc,” etc.) and multiple other rules to reduce the possible number of alias derivatives for a given company.

Finally, an important note is that we do not consider subsidiaries in determining co-occurrences. A main reason is that joint ownership of a subsidiary (by two or more companies) is frequent and may easily obscure the strength of a co-occurrence in news. For example, suppose 32% of Hulu is owned by NBCUniversal, which in turn is wholly owned

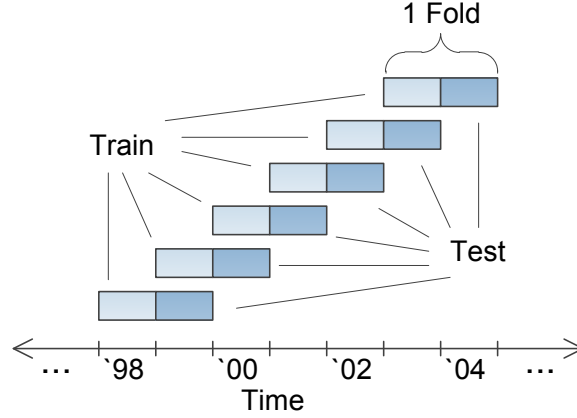


Figure 3.8: Walk-forward Testing

by Comcast. Further suppose there is a co-occurrence in news between Hulu and Google. What should be the strength of the relationship between Google and Comcast? What about the other owners of Hulu? We leave this for future work.

### 3.3 Predictiveness of Individual Datasets

We wish to use the similarity values in a predictive sense. Accordingly, we use walk-forward testing (Meyers, 1997; Aronson, 2007) as illustrated in Figure 3.8. This approach measures for a relationship between the predictor values (historical correlation, analyst cosine or news cosine) computed over one year's data and the stock return correlations in the subsequent year (i.e., the future correlation).

In Figures 3.9, 3.10 & 3.11, we group pairs of stocks into five separate ranges by their predictor values. Data from the entire S&P 1500 is used. The first observation that one might make is that correlation varies significantly between years and appears to be increasing in general. This trend is depicted by Figure 3.12 and has been observed by others (e.g., Tóth and Kertész (2006)). Further, as mentioned in section 2.2, it is well-known that correlations tend to increase in periods of crisis, which is evident in Figure 3.12 during two recent crash periods, 1987 (esp. Black Monday) and 2008 (the Global Financial Crisis).

For each of the predictors in Figures 3.9, 3.10 & 3.11, a higher range of values should mean higher future correlation, and this is true for each predictor with varying strength. For both historical correlation and analyst cosine values, a higher range of values nearly

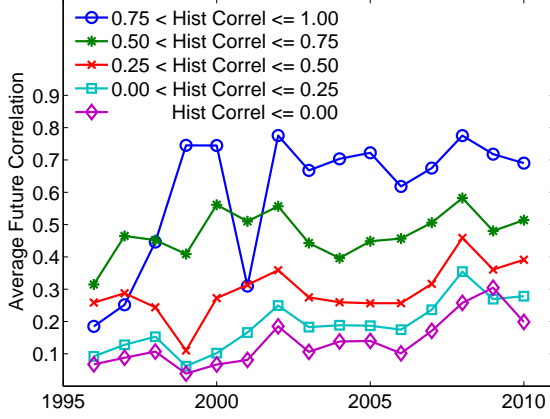


Figure 3.9: Hist. Correlation Predictiveness

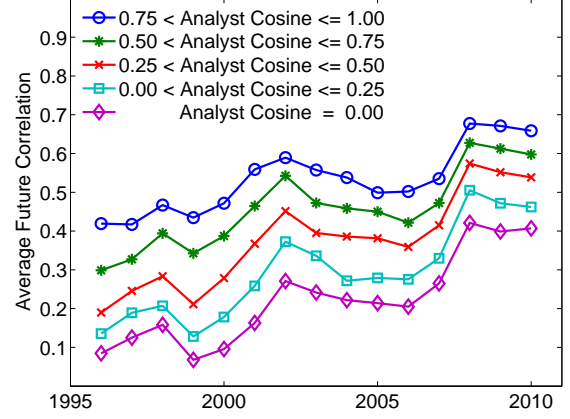


Figure 3.10: Analyst Predictiveness

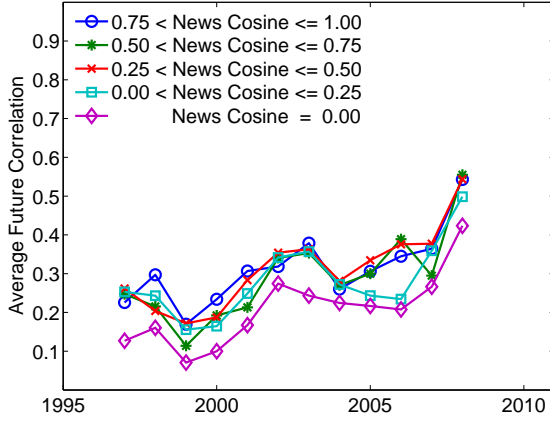


Figure 3.11: News Predictiveness

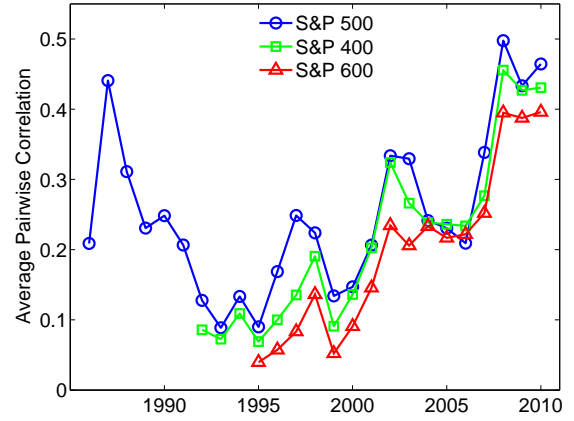


Figure 3.12: Correlation Time Dependence

always means higher future correlation. For news, the relationship is much noisier. Still the correlation for the highest range of news cosine values is higher than the correlation for the lowest range, indicating there is some predictiveness with news.

To test the statistical significance of the relationship between each predictor (historical correlation, analyst cosine and news cosine) and future correlation, we use the nonparametric Kendall's tau<sup>10</sup> (Kendall, 1938), which tests for a relationship by counting occurrences of concordance versus discordance. In this setting, concordance means that if we take two pairs of stocks, the pair with higher predictor value has higher future correlation. Discordance means the pair with higher predictor value has lower future correlation. Kendall's tau counts and normalizes all possible occurrences to output a number between

<sup>10</sup>We use the version of Kendall's Tau (called Tau-b) that accounts for ties.

Table 3.5: Kendall's Tau Values

Year	HIST. CORRELATION		ANALYST COSINE		NEWS COSINE	
	Kendall's Tau	Odds Ratio	Kendall's Tau	Odds Ratio	Kendall's Tau	Odds Ratio
1996	0.138	32.1%	0.094	20.6%	0.042	8.7%
1997	0.138	32.1%	0.108	24.1%	0.079	17.1%
1998	0.215	54.9%	0.110	24.7%	0.110	24.7%
1999	0.228	59.1%	0.086	18.9%	0.073	15.6%
2000	0.142	33.1%	0.108	24.3%	0.071	15.2%
2001	0.182	44.4%	0.120	27.2%	0.044	9.3%
2002	0.306	88.1%	0.111	24.9%	0.041	8.6%
2003	0.311	90.3%	0.106	23.6%	0.048	10.1%
2004	0.397	131.9%	0.102	22.6%	0.074	16.0%
2005	0.301	86.1%	0.083	18.1%	0.040	8.3%
2006	0.266	72.4%	0.083	18.0%	0.021	4.3%
2007	0.311	90.3%	0.084	18.3%	0.018	3.6%
2008	0.249	66.4%	0.069	14.9%	0.061	13.0%
2009	0.226	58.4%	0.091	20.0%	0.051	10.7%

$-1$  and  $1$ , where  $-1$  indicates all occurrences are discordant and  $1$  indicates all occurrences are concordant. Kendall's tau is also easy to interpret because an odds-ratio can be computed by  $(1 + \tau)/(1 - \tau)$ , where  $\tau$  is Kendall's tau. Thus, if  $\tau = 0.1$ , then the odds-ratio is  $(1 + 0.1)/(1 - 0.1) = 11/9 \approx 1.22$ , so concordance is 22% more likely than discordance. Table 3.5 shows the Kendall's tau values and corresponding odds-ratios when testing for a relationship between each predictor and future correlation. In each year and for each predictor, the Kendall's tau values are greater than zero with statistical significance well below 0.1%. So, each predictor is clearly positively related to future correlation. At the same time, it is clear that historical correlation has highest predictive strength, while analyst cosine has lower strength and, finally, news cosine is weakest. As will be seen in the next section, this does not mean that analysts and news are not useful.



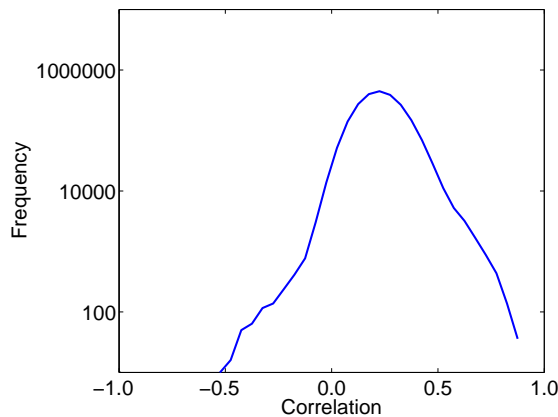


Figure 3.13: Historical Correlation Frequency Distribution

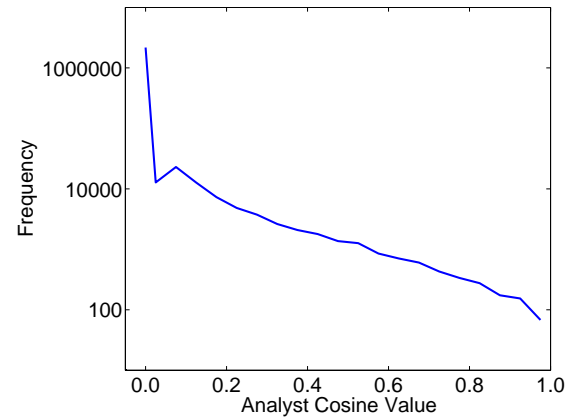


Figure 3.14: Analyst Cosine Frequency Distribution

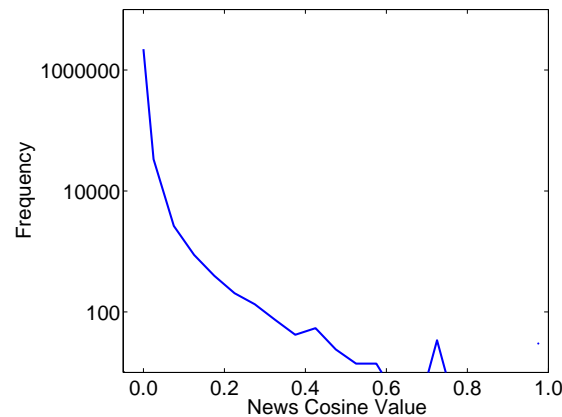


Figure 3.15: News Cosine Frequency Distribution

### 3.3.1 Selectivity Properties

In the previous section, Figures 3.9, 3.10 & 3.11 show promising results. Still, consideration must be given to the distribution of pairs of stocks that fall within each of the predictor value ranges in the figures. Consider Figures 3.13, 3.14 & 3.15. Correlation is characterized by a distribution that most closely follows a standard normal distribution compared to the others. For a pair of stocks selected at random, the most likely correlation values are near the mean value. Analyst and news cosines follow a much different distribution. Most values are near zero (i.e., there is no overlap in the sets for the two companies). High values are much less common. In fact, even though 99.6% of stocks in the S&P 1500 were covered by

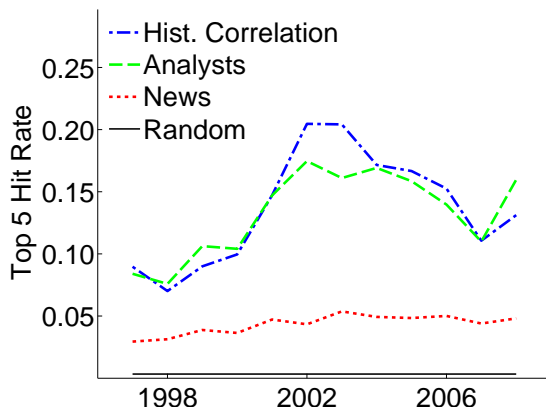


Figure 3.16: Top 5 Hit Rate

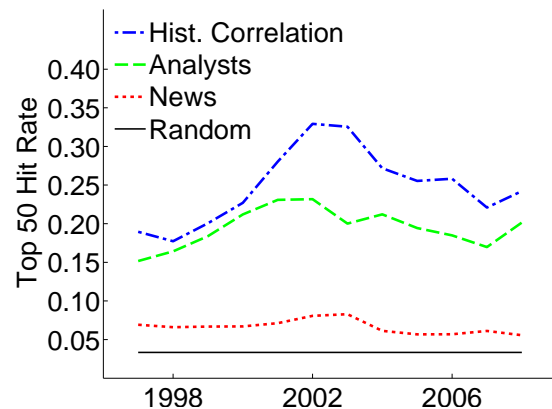


Figure 3.17: Top 50 Hit Rate

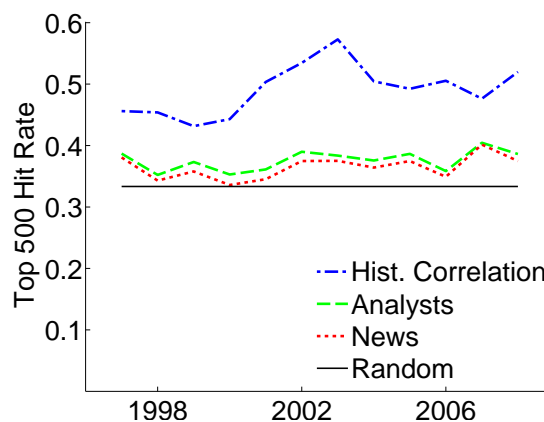


Figure 3.18: Top 500 Hit Rate

at least one analyst in 2010, roughly only 4% of pairs of stocks had any analyst covering both in the pair. The percentages are similar for all other years in the dataset. News has similar characteristics where a large majority of pairs of stocks never appear together in any articles. The end result is that news and analyst cosine values are generally not effective in differentiating moderately and slightly similar stocks, but are effective at determining highly similar stocks.

To demonstrate this phenomenon, we quantify the strength of each predictor with a “hit rate,” which we will explain through a hypothetical example. Suppose we wish to predict the top five most similar stocks to Walmart. We rank all other stocks using one of the predictors, such as the analyst cosine, on the previous year’s data. We then choose the top five most similar companies: Costco, Target, Sears Holdings, Macy’s and Rite Aid. Next, we determine the actual top five most correlated stocks for that year: Costco, Target,

Macy’s, Family Dollar and Safeway. Three out of the five stocks were predicted correctly, so our “hit rate” is  $3/5 = 0.6$ . Figures 3.16 , 3.17 & 3.18 display the average hit rate of all stocks in the S&P 1500.

As the figures indicate, the relative performance of the analyst cosine degrades from selecting the top 5 stocks to selecting the top 50 stocks and further degrades when selecting the top 500 stocks. This occurs because analysts will tend to cover relatively small sets of highly similar stocks. Moderately related stocks are not well differentiated from unrelated stocks because the analyst’s data presents little or no information on pairs of stocks that are not highly related. Figure 3.14 also supports this claim since most pairs of stocks have cosine of zero (i.e., both moderately related and unrelated pairs of stocks have cosine zero).

Also in the figure, it is evident that news has much less predictive power. However, its hit rates are far greater than random. The average top  $k$  hit rate for a random selection of stocks is

$$\sum_{j=1}^k \binom{j}{k} \frac{\binom{k}{j} \binom{N-k}{k-j}}{\binom{N}{k}} = \sum_{j=1}^k \binom{j}{k} h(j; N, k, k)$$

where  $h$  is the hypergeometric function and  $N$  is the number of stocks in the universe (1500 in our case for the S&P 1500). Thus, the average random top 5 hit rate is 0.0033, the average random top 50 hit rate is 0.0333 and the average random top 500 hit rate is 0.3333. For news, the minimum top 5 hit rate for any year is 0.0295, while the minimum top 50 hit rate is 0.0546. It is not until the top 500 that the minimum hit rate reaches 0.3358, nearly the same as random. These results indicate that news, just like analysts, is better at differentiating highly similar stocks than moderately or unrelated stocks. In chapter 6, we will see that determining the most similar stocks is important to many applications.

### 3.4 Method to Combine Datasets

Since each of the individual datasets, historical correlation, analysts and news, exhibit predictive power over future correlation, a natural next step is to combine these datasets in hope of obtaining better predictions. A naive approach would be to perform linear regression using previous data to determine weights, then using those weights for prediction.

For example,

$$\rho_{i,j,t} = \beta_0 + \beta_1 \rho_{i,j,t-1} + \beta_2 \mathcal{C}_{i,j,t-1}^A + \beta_3 \mathcal{C}_{i,j,t-1}^N + \epsilon_{i,j}$$

where  $\rho_{i,j,t}$  is the “current” correlation of stocks  $i$  and  $j$  at time  $t$ ,  $\rho_{i,j,t-1}$  is the “previous” correlation of  $i$  and  $j$  at time  $t - 1$ ,  $\mathcal{C}_{i,j,t-1}^A$  and  $\mathcal{C}_{i,j,t-1}^N$  are the cosines of analyst coverage and news co-occurrences of stocks, respectively, of stocks  $i$  and  $j$  using data from time  $t - 1$ . The regression coefficients are  $\beta_0$ ,  $\beta_1$ ,  $\beta_2$  and  $\beta_3$ , while the regression error is  $\epsilon_{i,j}$ . The purpose of this regression would be to produce these coefficients and apply them to the current year’s correlation ( $\rho_{i,j,t}$ ) and cosines ( $\mathcal{C}_{i,j,t}^A$  and  $\mathcal{C}_{i,j,t}^N$ ) such that a prediction can be made for the future correlation  $\rho_{i,j,t+1}$ .

However, if we consider again the dramatic year-over-year changes in correlation (as depicted in Figure 3.12), it is clear that the model will not perform well since it is predicting the absolute levels of correlation. Such prediction would require more sophistication and inclusion of more data, such as general economic indicators. For example, recall from section 2.2 that correlation tends to increase in periods of crisis, so predicting absolute levels of correlation would likely necessitate prediction of economic crises. Such prediction is beyond the scope of this work. Rather, we seek to know the similarity of pairs of stocks relative to other pairs of stocks. This relative similarity can be expected to remain more invariant from year to year than the absolute levels of correlation. At the same time, accurate prediction of relative levels of similarity is extremely useful, as will be shown in chapter 6.

To quantify relative similarity, we again use Kendall’s tau as was done in section 3.3 to quantify the predictiveness of the individual datasets. Recall that Kendall’s tau is a measure of concordance against discordance, where concordance occurs when a higher value of our predictor results in a higher value in future correlation and discordance occurs when a higher predictor value results in lower correlation. Here our predictor is composed of the individual predictors (historical correlation, analyst cosine and news cosine). To find the find weights for the individual predictors, we use Kendall’s tau as an objective function, seeking to find a combination of the datasets that maximizes Kendall’s tau using the previous year’s data. We then use that combination on the current year’s data to make a prediction for the subsequent year.

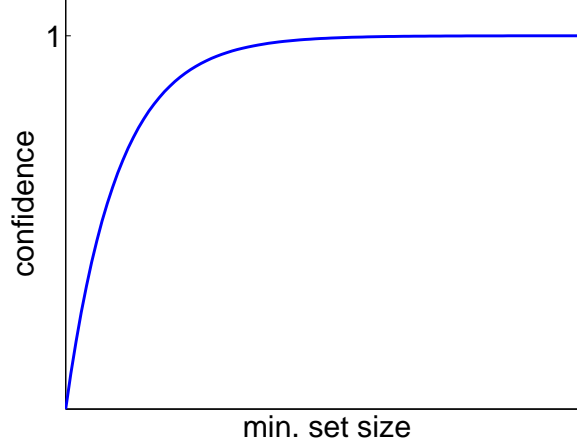


Figure 3.19: Confidence Coefficient Profile

### 3.4.1 Controlling for Confidence

Our approach to combining datasets is based on a simple linear combination of the predictors

$$\rho_{i,j,t-1} + w_A \cdot C_{i,j,t-1}^A + w_N \cdot C_{i,j,t-1}^N \quad (3.2)$$

where  $w_A$  and  $w_N$  are the weights of the analyst cosine and news cosine, respectively. Observe that this model has the favorable property that if the analyst and news cosine values are zero, the output value will simply be the historical correlation. That is, the analyst and news values will only add to the correlation if they are non-zero. This property fits with the observed distributions of these values in Figures 3.13, 3.14 & 3.15 and the observation that analysts and news are effective at selecting the most similar stocks. For less similar stocks, the output of the model “defaults” to using correlation.

We augment this model by incorporating a confidence multiplier for the analyst and news cosines. The confidence multiplier makes use of the intuition that the larger the sets used in the computation of the cosine, the higher confidence we can have in its value. For example, if two sets, each with ten elements, intersect on five elements, then their cosine is 0.5. Likewise, if two sets, each with one hundred elements, intersect on fifty elements, then their cosine is still 0.5. However, we can likely be more confident in the latter’s value since more elements were present in the computation. So, if more news stories are involved in the cosine computation, then it should have higher weight in the combination with correlation and analysts. Likewise, more analysts should translate to higher weight for the analyst

cosine value.

We compute the confidence multiplier for analysts cosine  $\mathcal{C}^A$  as

$$1 - e^{-\varphi_A \min(a_i, a_j)}$$

where  $a_i$  and  $a_j$  are the count of analysts covering stocks  $i$  and  $j$ , respectively. The rate of increase in confidence is controlled by  $\varphi_A$ , which will be optimized using the previous years data (as is done for  $w_A$  and  $w_B$  above). The minimum of the two set sizes is used in the exponent rather than some combination of their values because the size of the smaller set is most important to the sense of confidence. For example, if the sizes of the two sets are 10 and 1000, the addition of another element to the smaller set is much more important than another element added to the larger set. In the analyst dataset, such mismatches are frequent, and even more so with the news dataset. (See Figures 3.1 and 3.2.)

As shown in Figure 3.19, the functional profile for the confidence multiplier is that as the minimum of the two set sizes increases, the confidence increases. However, the greatest increases occur when the minimum set size is smaller. This matches intuition that suggests the marginal increase in confidence should be decreasing as the set sizes increase. For example, an increase from one to two analysts, should result in a larger increase in confidence than an increase from 100 to 101 analysts.

Similar to the analysts confidence multiplier, the confidence multiplier for news cosine  $\mathcal{C}^N$  is

$$1 - e^{-\varphi_N \min(m_i, m_j)}$$

where  $m_i$  and  $m_j$  are the count of news article co-occurrences for stocks  $i$  and  $j$ , respectively. The form is identical to the confidence computation for the analyst cosine, however the actual optimized value for  $\varphi_N$  may lead to a different rate of increase.

One might remark that the justification for using the cosine measure of similarity in section 3.1 was to control for differences in set sizes, and with this notion of confidence, we are again attempting to control for these same size differences. Importantly, the normalization used in the cosine measure simply helps to control for imbalances, but it does not provide any indication of the confidence. When combining with the other datasets, this measure of confidence is invaluable because it allows the model to provide more or less weight to the

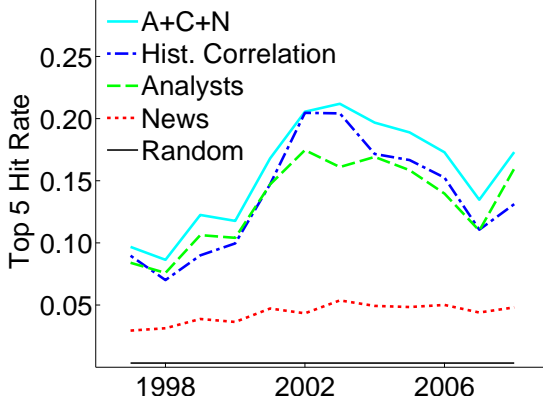


Figure 3.20: Top 5 Hit Rate

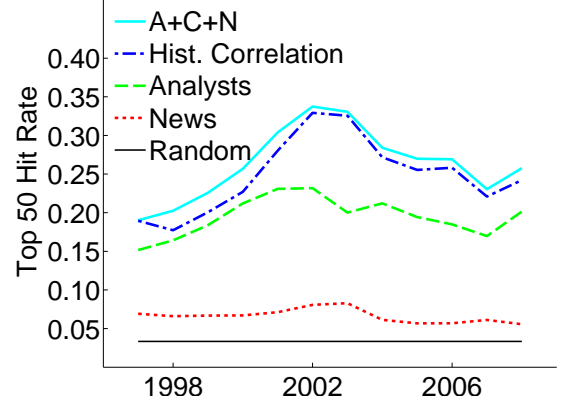


Figure 3.21: Top 50 Hit Rate

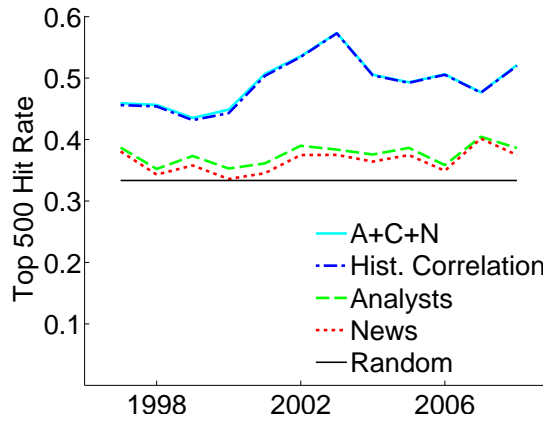


Figure 3.22: Top 500 Hit Rate

other predictors depending upon how much a given predictor should be trusted.

Incorporating these confidence values into the linear equation, the model becomes

$$\begin{aligned} \rho_{i,j,t-1} + w_A \cdot \left(1 - e^{-\varphi_A \min(a_{i,t-1}, a_{j,t-1})}\right) \cdot \mathcal{C}_{i,j,t-1}^A \\ + w_N \cdot \left(1 - e^{-\varphi_N \min(m_{i,t-1}, m_{j,t-1})}\right) \cdot \mathcal{C}_{i,j,t-1}^N \end{aligned} \quad (3.3)$$

To find optimal values for the coefficients ( $w_A$ ,  $w_N$ ,  $\varphi_A$  and  $\varphi_N$ ), we use Matlab's `patternsearch`<sup>11</sup>. For the objective, we implement a MEX-function that computes Kendall's tau in  $O(n \lg n)$  time based on the algorithm described by Knight (1966). This dramatically improves computation times over standard  $O(n^2)$  approaches. After these optimal weights are computed for each year, they are used to make predictions for the subsequent year.

Results are shown in Figures 3.20, 3.21 and 3.22, which reproduce Figures 3.16, 3.17 and

<sup>11</sup><http://www.mathworks.com/help/gads/patternsearch.html>

3.18, but with the confidence-based combination of analyst, correlation and news (A+C+N) also included. In the figures, there are clear improvements in the hit rates for the most similar stocks (i.e., top 5 and top 50). Using a one-tailed paired t-test of the hit rates over the years, the top 5 hit rate for A+C+N is significantly higher than both historical correlation ( $p = 3.8 \times 10^{-5}$ ) and analysts ( $p = 1.1 \times 10^{-5}$ ). The top 50 hit rate is also significantly higher than historical correlation ( $p = 6.4 \times 10^{-5}$ ) and analysts ( $p = 2.1 \times 10^{-6}$ ). Importantly, performance is generally no worse than the best predictor, so including the other factors does not cause harm. This is evident even in the top 500 hit rate, where performance of the combination method does not drop below historical correlation alone.

### 3.4.2 Importance of Factors

The A+C+N model (equation 3.3) has multiple inputs and therefore we consider the importance of each of these inputs in making predictions.

#### Predictors

To evaluate the value of analysts, correlation and news, we consider models that combine only two of the three predictors:

**correlation + analysts:**

$$\rho_{i,j,t-1} + w_A \cdot (1 - e^{-\varphi_A \min(a_{i,t-1}, a_{j,t-1})}) \cdot \mathcal{C}_{i,j,t-1}^A$$

**correlation + news:**

$$\rho_{i,j,t-1} + w_N \cdot (1 - e^{-\varphi_N \min(m_{i,t-1}, m_{j,t-1})}) \cdot \mathcal{C}_{i,j,t-1}^N$$

**analysts + news:**

$$(1 - e^{-\varphi_A \min(a_{i,t-1}, a_{j,t-1})}) \cdot \mathcal{C}_{i,j,t-1}^A + w_N \cdot (1 - e^{-\varphi_N \min(m_{i,t-1}, m_{j,t-1})}) \cdot \mathcal{C}_{i,j,t-1}^N$$

We abbreviate these models as A+C, C+N and A+N, respectively.

In Table 3.6, we display the Kendall's tau values between correlation and the outputs of the combinations based on the previous years data. Recall that Kendall's tau is the objective measure in our parameter optimization process, but it is also a global measure. So, whereas we are most interested in the top K results, Kendall's tau is a measure across all



Table 3.6: Kendall's Tau Values for Confidence Augmented Combinations

year	Ana	Cor	News	A+C	A+N	C+N	A+C+N
1997	0.1097	0.2153	0.1100	0.2199	0.1376	0.2171	<b>0.2215</b>
1998	0.0861	0.2281	0.0725	0.2286	0.1012	0.2284	<b>0.2291</b>
1999	0.1081	0.1421	0.0707	0.1454	0.1190	0.1430	<b>0.1458</b>
2000	0.1197	0.1818	0.0444	0.1869	0.1184	0.1821	<b>0.1870</b>
2001	0.1109	0.3057	0.0412	0.3069	0.1109	0.3060	<b>0.3071</b>
2002	0.1056	0.3110	0.0482	0.3117	0.1073	0.3112	<b>0.3119</b>
2003	0.1017	0.3974	0.0740	0.3976	0.1185	0.3981	<b>0.3982</b>
2004	0.0829	0.3009	0.0401	<b>0.3020</b>	0.0853	0.3004	0.3014
2005	0.0827	0.2659	0.0210	<b>0.2663</b>	0.0751	0.2659	<b>0.2663</b>
2006	0.0837	0.3109	0.0179	0.3114	0.0837	0.3110	<b>0.3114</b>
2007	0.0693	0.2492	0.0610	0.2495	0.0693	0.2495	<b>0.2497</b>
2008	0.0908	0.2259	0.0506	0.2268	0.0979	0.2261	<b>0.2268</b>

A+N and A+C+N are tied in 2005 because the optimizer returned a zero weight for news ( $w_N$ ) using the 2004 data.

pairs in the stock universe. Small differences in Kendall's tau may mean large improvements in top K results.

From Table 3.6 it can be seen that historical correlation is the best predictor globally, followed by analysts then news. This same result was already discussed in section 3.3. However, if we look at the combinations of two predictors (A+C, A+N and C+N), it is evident that each combination does better than either of its single predictors alone. Using paired t-tests over the years, A+C outperforms historical correlation ( $p = 0.0045$ ), A+N outperforms analysts ( $p = 0.029$ ) and C+N outperforms historical correlation ( $p = 0.022$ ). Furthermore, A+C+N outperforms A+C ( $p = 0.049$ ). Although it is evident that news brings least improvement, each of these results is statistically significant at the 5% level.

Table 3.7: Kendall’s Tau Values for Simple Linear Combinations

year	Ana	Cor	News	A+C	A+N	C+N	A+C+N
1998	0.0861	0.2281	0.0725	<b>0.2292</b>	0.1018	0.2278	0.2289
1999	0.1081	0.1421	0.0707	0.1449	0.1192	0.1424	<b>0.1450</b>
2000	0.1197	0.1818	0.0444	<b>0.1861</b>	0.1175	0.1804	0.1856
2001	0.1109	0.3057	0.0412	<b>0.3063</b>	0.1109	0.3057	0.3063
2002	0.1056	0.3110	0.0482	0.3114	0.1056	0.3112	<b>0.3115</b>
2003	0.1017	0.3974	0.0740	0.3976	0.1185	0.3976	<b>0.3978</b>
2004	0.0829	0.3009	0.0401	<b>0.3020</b>	0.0851	0.2998	0.3008
2005	0.0827	0.2659	0.0210	<b>0.2663</b>	0.0749	0.2659	0.2663
2006	0.0837	0.3109	0.0179	0.3114	0.0837	0.3110	<b>0.3114</b>
2007	0.0693	0.2492	0.0610	0.2494	0.0693	0.2493	<b>0.2495</b>
2008	0.0908	0.2259	0.0506	<b>0.2266</b>	0.0976	0.2254	0.2260

### Confidence

Finally, we evaluate the importance of incorporating the confidence multipliers for analysts and news. To make this assessment, we examine using only the simple linear combination of equation 3.2. We also consider the simple linear combinations of two predictors:

$$\text{correlation} + \text{analysts: } \rho_{i,j,t-1} + w_A \cdot \mathcal{C}_{i,j,t-1}^A$$

$$\text{correlation} + \text{news: } \rho_{i,j,t-1} + w_N \cdot \mathcal{C}_{i,j,t-1}^N$$

$$\text{analysts} + \text{news: } \mathcal{C}_{i,j,t-1}^A + w_N \cdot \mathcal{C}_{i,j,t-1}^N$$

Results are shown in Table 3.7. Against the confidence-based combinations (Table 3.6), all simple combinations perform worse, although only C+N and A+C+N are significantly worse ( $p = 0.011$  and  $p = 0.024$ , respectively). The performance reduction for A+C is not significant ( $p = 0.105$ ) and neither is A+N ( $p = 0.442$ ). Regardless, whereas the confidence-based C+N is significantly better than correlation alone, the simple linear form of C+N actually performs worse than correlation alone. The same is true when comparing A+C+N with A+C — the addition of news does not bring any benefit in a simple linear combination. This implies that the confidence multiplier is extremely important when including news in

the combination.

### 3.5 Summary

- Analyst cosine and news cosine are found to be predictive of future correlation, particularly for the most similar pairs of companies. That is, analysts and news perform best for the “top K” pairs of stocks where K is a small number. As K increases, performance degrades. Historical correlation is also shown to be predictive of future correlation, but for all pairs of stocks (i.e., regardless of K). These performance profiles for analysts and news and for historical correlation can be explained by their fundamental properties. On one hand, historical correlation can be computed for any two timeseries of returns, so historical correlation can produce a similarity value for any pair of stocks. On the other hand, analysts tend to cover highly similar companies and, likewise, similar companies are much more likely to appear together in news. So, analysts and news are not good at differentiating between companies with moderate or little similarity. For example, analysts and news are good determining if Microsoft is more similar to Google than to Apple, but not if Microsoft is more similar to Ford than to McDonald’s. However, the most similar stocks (i.e., low K in “top K”) are important for many applications, such as hedging and relative valuation.
- Methods to combine the analyst, correlation and news data (A+C+N) are considered. A linear combination is used with a confidence multiplier for the analyst and news cosines that gives greater weight to the cosine values when the companies are covered by more analysts or have more occurrences in news articles. The end result is that the A+C+N combination is found to better predict future correlation than any of its individual inputs, particularly for the most similar pairs of companies (i.e., the “top K”).

## Chapter 4

### Hypergraph Partitioning

Chapter 3 focused on predicting the similarity of a pair of stocks, specifically predicting the future correlation of returns for those stocks. In chapter 5, the focus turns to forming groups of highly similar stocks, analogous to the industry taxonomies described in section 2.3. As will be seen in the next chapter, hypergraph partitioning is an intuitive method that can be used to form groups from the analyst and news data. Each vertex in the hypergraph corresponds to a single company. With the analyst data, each edge (i.e., hyperedge) corresponds to a single analyst and connects the companies that analyst covers. With the news data, each edge corresponds to a news article and connects the companies in the article. Using hypergraph partitioning, the companies will be split into parts such that the number of cut hyperedges is essentially minimized. More details on performance will be discussed in chapter 5, but the purpose of this chapter is to describe a hypergraph partitioning algorithm that respects an entropy constraint. In forming stock groups, this constraint is important because it enables fair comparisons with industry taxonomies. More generally, there are many application areas where such a constraint is desirable, such as consensus clustering (i.e., cluster ensembles). This chapter is devoted to the presentation of the algorithm and comparison with existing algorithms. Datasets outside the financial domain are used in order to illustrate the generality of the algorithm.

#### 4.1 Motivation

A hypergraph is similar to a graph, except edges can connect any number of vertices, whereas edges in a graph each connect exactly two vertices. Formally, a hypergraph  $H = (V, E)$  where  $V$  is a set of vertices and  $E$  is a set of edges such that for all  $e \in E, e \subseteq V$ . Edges are sometimes called hyperedges, nets or links. The vertices that compose an edge are called

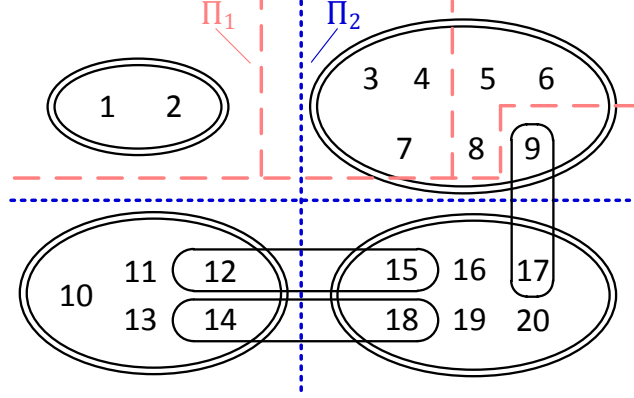


Figure 4.1: Hypergraph Partition Example

its pins.

The partition problem involves splitting the vertices  $V$  into disjoint sets  $(V_1, V_2, \dots, V_k)$  such that the number of edges with pins appearing in more than one part is minimized (i.e., minimize the edge “cut”). Left unconstrained, solving the partition problem can often lead a degenerate solution with several tiny parts that simply have the fewest edges connecting their vertices. To ensure reasonable sizes, a balance constraint is typically imposed. Since partitioning a graph under balance constraints is NP-Hard (Garey, 1979), the same is true for hypergraph partitioning.

A major application area of hypergraph partitioning is integrated circuit (IC) design (Papa and Markov, 2007), where circuits are divided such that connections between subdivisions are minimized. In this setting, a common balance constraint (Karypis et al., 1997) is

$$\frac{1}{c} \cdot \frac{|V|}{k} \leq |V_i| \leq c \cdot \frac{|V|}{k} \quad (4.1)$$

where the imbalance tolerance is specified by  $c \geq 1.0$ . This matches an IC specification that parts be roughly equal size.

Yet, this constraint is not necessarily appropriate for all domains, especially those where natural clusters are “quasi-balanced.” Consider the hypergraph in Figure 4.1, where edges are indicated by black ovals. Vertex sets  $\{1, 2\}$ ,  $\{3, 4, 5, 6, 7, 8, 9\}$ ,  $\{10, 11, 12, 13, 14\}$  and  $\{15, 16, 17, 18, 19, 20\}$  are each enclosed by two edges, and make a reasonable 4-way partition  $\Pi_2$ , as indicated by the short-dashed blue line. In order to allow the small set  $\{1, 2\}$  to be

its own part,  $c = |V|/(2k) = 20/(2 \cdot 4) = 2.5$ . However, this also allows the largest part to be bounded by  $c|V|/k = 2.5 \cdot 20/4 = 12.5$ . Thus, the partition  $\Pi_1$  is allowed and has a lower cut of 2, instead of a cut of 3 for  $\Pi_2$ .

One might consider replacing  $c$  of equation 4.1 with separate parameters for the upper and lower bounds. Such alterations could prevent the situation of Figure 4.1, but do not solve the underlying issue in that the notion of “imbalance” is determined solely by the part with largest deviation from the average size, while deviations of other parts offer no contribution. For example, in a 6-way partition of 60 vertices, these sets of part sizes would each have equal imbalance under equation 4.1 or similar formulations:  $\{2,10,10,10,10,18\}$ ,  $\{2,4,8,12,16,18\}$  and  $\{2,2,2,18,18,18\}$ . We posit the first set is more “balanced” than the second, and even more so than the third. Meanwhile, we suggest that many quasi-balanced datasets may have natural clusters with sizes similar to any of the three.

We suggest that a more appropriate balance constraint is information-theoretic entropy

$$c_\ell \leq - \sum_i^k \frac{|V_i|}{|V|} \lg \left( \frac{|V_i|}{|V|} \right) \leq c_u \quad (4.2)$$

where  $c_\ell$  and  $c_u$  are user-specified constants defining the range of acceptable solutions. In Figure 4.1,  $\Pi_1$  is far less balanced in the information theoretic sense, with entropy 1.596, whereas  $\Pi_2$  has entropy 1.883. Thus, the desired partition would be achieved by setting  $c_\ell$  anywhere between 1.596 and 1.883 and  $c_u$  beyond 1.883, up to the maximum possible value of 2 corresponding to purely equal-sized parts of 5. The partition algorithm is free to choose small or large parts as long as the overall balance stays within the constraint. This contrasts with typical constraints, such as formula 4.1, where the user is forced into an all-or-none situation when setting  $c$ , leaving the partition algorithm free to push *all* part sizes to the extremes.

Hypergraph partitioning has recently been employed to perform consensus clustering (a.k.a. cluster ensembles), which is the process of combining the results of clustering algorithms or runs of the same algorithm to compute a single clustering. Perhaps because most tools and research on hypergraph partitioning have focused on equal-sized partitions, its use has been limited to problems where roughly equal-sized clusters are desired. In fact, a recent survey (Ghosh and Acharya, 2011) suggests that “employing a graph clustering

algorithm adds a constraint that favors clusterings of comparable size.” We do not believe this limitation is necessary.

Building on existing techniques, we present an algorithm *hyperpart* that respects the entropy constraint of formula 4.2. Additionally, we develop a new cut cost measure, *discount cut*, that helps avoid local minima, a known weakness in many k-way partitioning algorithms (Cong and Lim, 1998). Comparing to today’s leading partitioners, we demonstrate our algorithm is best able to produce high quality partitions for imbalanced datasets. We further show that by use of our algorithm, hypergraph partitioning can be effective in consensus clustering, even when cluster sizes are not roughly equal.

## 4.2 Background and Related Work

### 4.2.1 Progression of Partitioning Algorithms

Major advances in graph partitioning begin with the Kernighan-Lin algorithm (Kernighan and Lin, 1970), which makes vertex swaps in an equal-sized bi-partition to reduce its cut. The Fiduccia-Mattheyses (FM) algorithm (Fiduccia and Mattheyses, 1982) followed with the introduction of single-vertex moves and a balance constraint. In its original form, FM only performs bi-partitioning, but it can be applied recursively it to achieve k-way partitioning. Sanchis (1989) extended the FM algorithm to directly achieve a k-way partition with “k-FM”. Cong and Lim (1998) noted that k-FM is easily susceptible to local optima in practice and introduced K-PM/LR, which performs FM for pairs of parts in parallel, leading to large improvements in cut size.

Karypis and Kumar (1996) introduced the multi-level framework for partitioning, which consists of three phases. First, the hypergraph is coarsened by merging vertices and/or edges through a variety of heuristics. This process of consolidation may occur multiple times, with each called a *level*. Once the hypergraph is considered small enough, it is partitioned randomly or through some other heuristic. FM and/or other refinement algorithms are then applied until no improvements can be observed. The levels of coarsening are then unwound, with part assignments for each merged vertex in the coarsened graph applied to each of its corresponding de-merged vertices. Refinement algorithms are then re-applied.

The uncoarsening continues until the original graph is recovered. This framework has been shown to dramatically improve execution time, while also improving partition quality as measured through cut size. This framework has been applied to both recursive bi-partitioning (Karypis et al., 1997) and direct k-way partitioning (Karypis and Kumar, 2000). One drawback is the difficulty in enforcing a balance constraint during the coarsening and uncoarsening phases, particularly as more imbalance is allowed. As our experimental results will show, some partitioners are unable to produce any results at high levels of imbalance.

For purposes of comparison in this work, we consider two of the most popular recursive bi-partitioners, PaToH (Çatalyürek and Aykanat, 1999) and hMETIS (Karypis and Kumar, 1998). We also consider hMETIS’s k-way counterpart, khMETIS. (For simplicity, we refer to PaToH, hMETIS and khMETIS as *patoh*, *hmetis-rb* and *hmetis-kway*, respectively.) All three use a multi-level framework.

#### 4.2.2 Quality Measures & Balance Constraints

A common partition objective is to minimize the edge cut, which simply counts the edges that span more than one part. More sophisticated partitioners may minimize the Sum of External Degrees (SOED), which assigns a penalty equal to the number of parts an edge spans when it is cut. The similar K-1 measure has penalty of parts spanned minus one. Both emphasize reducing the number of parts spanned. For example, an edge with pins in 10 parts is less desirable than having pins in 2 parts. Both scenarios have cut 1, but SOED of 10 and 2, respectively, and K-1 of 9 and 1.

Balance constraints are typically defined in a manner that fits the algorithm. For the recursive bi-partitioner *hmetis-rb*, balance is controlled through the *UBFactor* denoted  $b$ . For each bisection, *hmetis-rb* will produce parts with sizes  $(50 - b)n/100$  and  $(50 + b)n/100$ , where  $n$  is the number of vertices. Using the example from the manual (Karypis and Kumar, 1998), with  $b = 5$  the parts will be between  $0.45n$  and  $0.55n$ . Further, in a 4-way partition, this means parts will be between  $0.45^2n = 0.2n$  and  $0.55^2n = 0.3n$ .

No longer focusing on bisection, *hmetis-kway* redefines the *UBFactor* such that “the heaviest [part] should not be  $b\%$  more than the average weight.” So, if  $b = 8$  and  $k = 5$ , the largest part(s) will have at most  $1.08n/5$  vertices. (Karypis and Kumar, 1998)



While still a recursive bi-partitioner, patoh defines its *imbal* parameter similarly to hmetis-kway, which is a threshold tolerance above average for the maximum sized part. Patoh internally adjusts its allowed imbalance during each bisection to meet the final k-way imbalance threshold. (Çatalyürek and Aykanat, 1999)

Note that each of these balance constraint definitions are essentially tolerances above (and sometimes below) an average part size. As suggested in section 4.1, with this approach imbalance is defined solely by the part with largest deviation from the average, while deviations of other parts are not considered. In contrast, our constraint (formula 4.2) relies on entropy, which is an information-theoretic measure for the uncertainty of a random variable. In the case of partitioning, if we choose a vertex uniformly at random, entropy can be used to measure the uncertainty about which part the vertex is assigned. Suppose we have two partitions,  $\Pi_a$  and  $\Pi_b$ , with sizes  $\{1,7,7,7\}$  and  $\{4,5,5,5\}$ , respectively. With  $\Pi_a$ , there is more certainty because part 1 is very unlikely, whereas  $\Pi_b$  has less certainty since all parts are closer to the same likelihood. In fact, the entropies of  $\Pi_a$  and  $\Pi_b$  are 1.287 and 1.577, respectively. Entropy is commonly used to compare the imbalance of clusterings (for example, see Meilă (2007)). Thus, we find it a more natural and more precise means to control imbalance.

Some partitioners allow the user to specify desired sizes per part. For example, k-FM uses constraints

$$r_i \cdot |V| - \epsilon \leq |V_i| \leq r_i \cdot |V| + \epsilon \quad (4.3)$$

for all  $1 \leq i \leq k$ , where  $V_i$  denotes part  $i$ ,  $0 \leq r_i \leq 1$ ,  $\sum_{i=1}^k r_i = 1$  and  $\epsilon$  denotes the error tolerance. These “target size” constraints are useful when parts of varying *fixed* sizes are desired, as in IC applications where die sizes are known, or in computational load balancing when resources have varying capacities. However, if specific part sizes are not known a priori, these constraints are of little use other than perhaps allowing the user to perform multiple runs over many sets of varying part sizes.

Others have sought “natural” partitions, especially by incorporating cut costs and balance constraints into a single objective. In a seminal work, Wei and Cheng (1989) present

ratio-cut minimization

$$\frac{C_{\pi_1 \pi_2}}{|\pi_1| \times |\pi_2|} \quad (4.4)$$

where  $\pi_1$  and  $\pi_2$  denote the parts after bisection and  $C_{\pi_1 \pi_2}$  denotes the edges cut by that partition. A lower cut size is favored by the numerator, while the denominator favors equal-sized parts. It can also be viewed as actual cut divided by expected cut. Thus, the ratio-cut optimizes a tradeoff between minimizing cut and maintaining balance, which can lead to more natural partitions where edges cut are lowest in comparison to expectation. Yet, it does not satisfy a situation where the user has some tolerance for imbalance and simply wishes to minimize the cut within that tolerance. For example, a 20-30 partition of a 50 vertex hypergraph with cut 100 would be preferred by the ratio cut to a 15-35 partition with cut 90, but the user may have been more satisfied with the latter partition simply because cut is lower.

The ratio cut has been extended to k-way partitioning with objectives such as scaled cost (Chan et al., 1994) and numerous other objectives have been proposed to address specific situations or applications (see Alpert and Kahng (1995) for a survey). Still, we suggest that minimizing the cut within an entropy constraint is most effective when equal parts are not required, fixed part sizes are unknown a priori and the user does not want to be forced into specific tradeoffs between imbalance and edge cut.

#### 4.2.3 Consensus Clustering

The process of combining the results of several cluster algorithms (or multiple runs of the same algorithm) to form a single clustering is known as consensus clustering or cluster ensembles. The motivation is that the consensus is often more robust than individual cluster algorithms. A simple, intuitive consensus clustering algorithm is the hypergraph partitioning algorithm (HPGA) (Strehl and Ghosh, 2003). Each cluster from each of the underlying cluster algorithms is represented by a hyperedge. The task is simply to partition the hypergraph such that the cut is minimized. Since only tools (esp. hmetis-rb) with equal part constraints have been used thus far, HPGA’s successful application has been limited to problems where clusters are to be roughly equal size. In a study comparing consensus clustering algorithms, Topchy et al. (2005) remarked “HPGA did not work well due to

its bias toward balanced cluster sizes.” We believe this is a limitation of the tools, not hypergraph partitioning itself. Thus, we view this work as a means to broaden the class of problems to which HPGA can successfully be applied.

While HPGA is not necessarily the most effective consensus clusterer, it is easy to understand and frequently appears as a baseline. Moreover, other algorithms, such as the Meta-Clustering Algorithm (MCLA) (Strehl and Ghosh, 2003) also use hypergraph partitioning as a component of its overall algorithm. We do not consider MCLA or other algorithms in this work, but believe use of an entropy constraint may have benefits beyond our focus on HPGA.

### 4.3 Method

#### 4.3.1 Algorithm Outline

Our partitioner *hyperpart* focuses on making single vertex moves that best improve the cut measure while still respecting the entropy constraint in formula 4.2. This improvement in cut is known as the move’s *gain*. After each vertex is moved, it is locked for the remainder of the *pass*, which concludes when all vertices are moved. Vertex locking, along with forcing all vertices to move even if the move has negative gain, serves as a form of local optima avoidance and also helps prevent oscillations. At the end of each pass, the moves are unwound back to the configuration where the cut was lowest, so negative moves are not necessarily kept until the next pass. These passes continue until no improvement is witnessed during an entire pass. Figure 4.2 depicts this general approach. The concepts of passes and vertex locking first appeared in the Kernighan & Lin algorithm, while the concept of move gains first appeared in the FM algorithm.

#### 4.3.2 Discount Cut

In finance, the “time value of money” describes the principal that, due to inflation, a dollar today can buy more than a dollar in the future. To estimate the value of a payment to be received in the future, one should discount it by the inflation rate. We use this notion to help overcome the well-known issue (Cong and Lim, 1998) of convergence to local optima

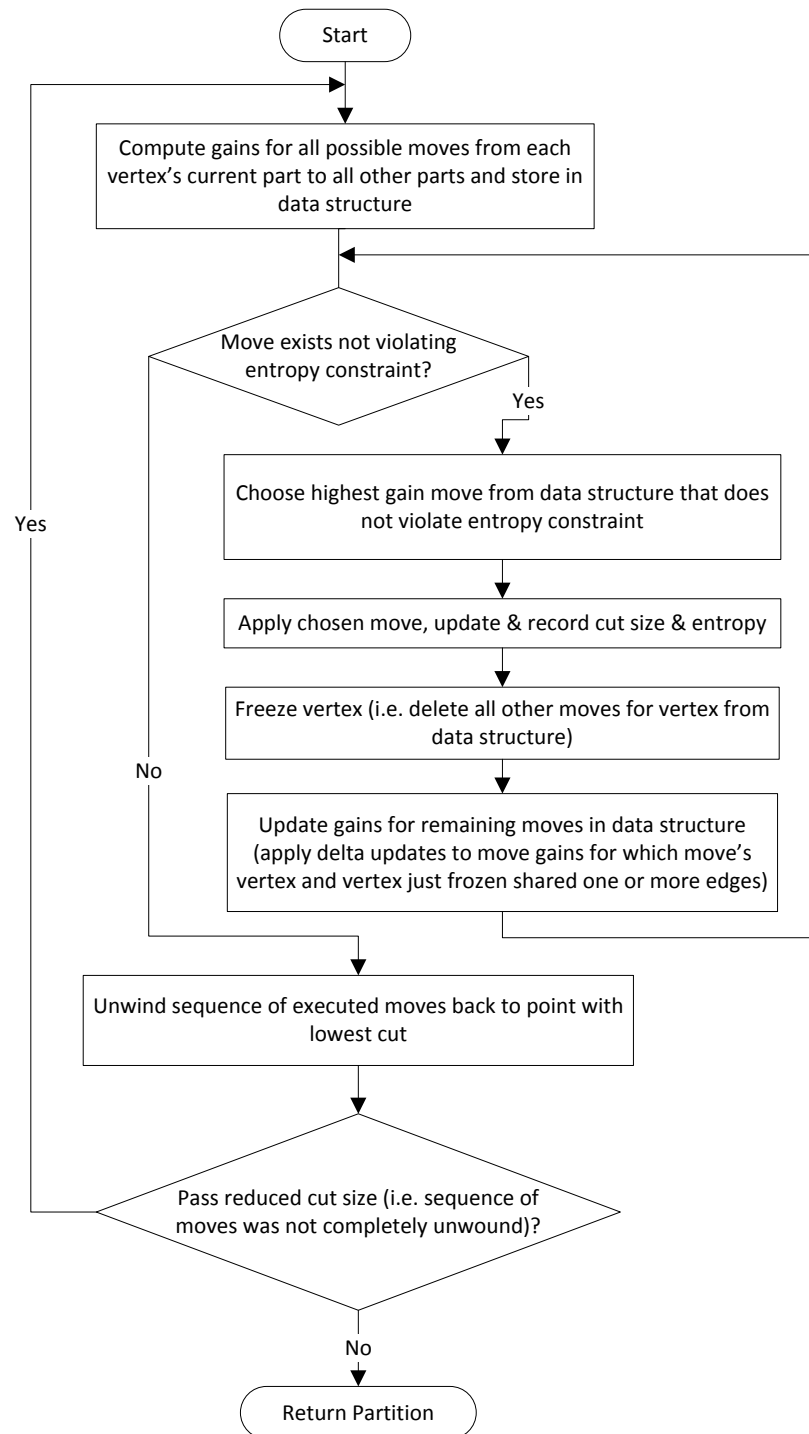


Figure 4.2: Hyperpart Flowchart

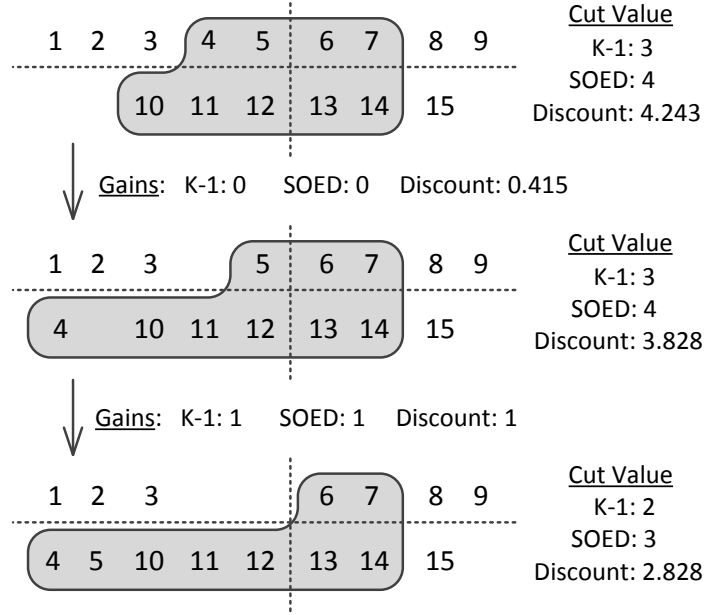


Figure 4.3: Discount Cut. Three different configurations of a single edge are depicted as shown by the grey areas. Vertices are represented by numbers, which are split by the dotted lines to show the partition (with 4 parts). Only one edge is depicted, but the full graph would normally have many edges. Cut values under different measures are shown right of each configuration. Gains for each measure are shown right of arrows, which indicate a vertex move leading from one configuration to the other.

for k-way FM-based partitioners. We define the discount cut as

$$\sum_{e \in E} \left[ \sum_{\pi \in \Pi, \pi \neq \hat{\pi}_e} |\pi \cap e|^\alpha \right] \quad (4.5)$$

where  $E$  is the set of edges,  $\Pi$  is the partition and  $0 \leq \alpha \leq 1$ . The part containing the most pins in edge  $e$  is denoted by  $\hat{\pi}_e$ . Only one part may be  $\hat{\pi}_e$ , so ties can be settled with a rule, such as choosing the part with lower label. Higher  $\alpha$  means future moves are more strongly discounted (i.e., worth less). In our experiments, we set  $\alpha = 0.5$ .

Consider Figure 4.3. The bottom configuration is most desirable because the edge spans only three parts. Two moves are required to get from top to bottom. Yet, if only the K-1 (or SOED) measure is used, the partitioner cannot see any benefit in moving from the top to the middle. These measures do not facilitate a view into the future, and it becomes easy for a partitioner to become stuck at a local optima. The concepts of passes and vertex

freezing may help to avoid this problem since vertices are forced to move, even if there are no positive gain moves for that vertex. Yet, there is still no guidance as to which moves may be beneficial, so making the best move is a matter of chance. For example, moving vertex 4 from the upper left to the lower left as shown in the diagram has the same gain of zero (the best gain possible) as moving vertex 12 from the lower left to lower right part. With the discount cut, there is positive gain in moving vertex 4. The discount cut for the top configuration is  $2^{0.5} + 2^{0.5} + 2^{0.5} = 4.243$ , while the discount cut for the middle is  $1^{0.5} + 2^{0.5} + 2^{0.5} = 3.828$ . So, there is a gain of  $4.243 - 3.828 = 0.415$ , and the partitioner is encouraged to make this move. In general, the discount cut rewards moves that take an edge's pin from one of its smaller parts to a larger part. The greatest reward is given to moving to the part with most pins since this part offers no contribution to the cut cost (as seen in formula 4.5).

As edge size increases, the significance of the discount cut in local minima avoidance grows since more pins means more moves will be necessary to realize gains in the K-1 measure (see section 4.4.2 for experimental evidence). At the same time, the best partition as computed by minimizing the discount cut may not exactly match the minimum K-1 partition, although they will likely be similar. Therefore, we first run one iteration of the algorithm using the discount cut cost. We then take the part assignments and repeat the algorithm using the K-1 measure to match the true objective of minimizing the number of parts each edge spans. Finally, as further local optima avoidance, we use random restarts with the initial part assignments reshuffled and with each part receiving an approximately equal number of vertices.

### 4.3.3 Implementation & Complexity

Hyperpart is implemented in C++ using STL and Boost libraries. It's input file is compatible with hmetis, where vertices and edges are numbered  $[1...|V|]$  and  $[1...|E|]$ , respectively. Code and binaries have been made available online at <http://john.robert.yaros.us/software>. Upon initialization, hyperpart constructs two arrays of bucket-lists. The first maps each vertex to its associated edges, while the second maps each edge to its associated vertices. The time to populate these arrays is  $O(|P|)$  since it depends on the number of

times each vertex appears in an edge, which is exactly the number of pins. The number pins in each part is also maintained for each edge and similarly requires  $O(|P|)$  to initialize. We maintain a each vertex's current part assignment in an array, which is initialized (with random assignments) in  $O(|V|)$  time and is easily updated in  $O(1)$  time as moves are executed. Recall that computation of the discount cut requires knowledge of each edge's max part ( $\hat{\pi}_e$  in formula 4.5). To track the max, we use a heap, which will have size equal to the number of parts  $k$ , so the total initialization time for all heaps is  $O(|P| \lg k)$ .

An essential key to fast execution is the ability efficiently maintain gains associated with each possible move and to quickly determine the move with highest gain. To do so, we use a binary search tree of gains, where each gain has a pointer to a linked list of moves with that particular gain. The maximum number of moves possible is  $(k-1)|V|$ , since any vertex can move to any part other than its current part. The tree will contain at most  $(k-1)|V|$  nodes, leading to  $O(\lg k|V|)$  insertion and deletion time. No additional time complexity is needed to maintain the lists at each node because we maintain with each move object a pointer to its linked list location (among the moves of same gain), so deletion from the list can be performed in constant time. Insertion simply involves appending to list, which is also constant time. To fully initialize this the tree will take  $O(k|P|)$  insertions since, for each vertex we must consider the cost of moving to each of the other  $k-1$  parts, and, for each possible part, one must consider the change in cut to each of the vertex's edges. However, one may observe from K-1 cut and discount cut formulas that determining the change in the cut measure does not actually require accessing the number of pins in all parts. The cut "delta" simply requires knowing the sizes of the origin and destination parts. One exception for the discount cut is that the edge's max part may change as a result of the move. This means that the sizes of the max and second max (in case the max is the origin part) must also be known, but these can be obtained in  $O(1)$  time from edge's part size heap. So, only two part sizes must be known for the K-1 measure and four for discount measure (not  $k$ ). The map's total initialization time is therefore  $O(k|P| \lg k|V|)$ . Across these data structures, we can see that the gain tree has dominate initialization time, so our the complexity for initialization is  $O(k|P| \lg k|V|)$ .

After initialization, the algorithm begins a pass whereby moves with highest gain are

executed one at time with each vertex locked until all vertices are locked. After each executed move, the gains of other moves must be updated, which is an  $O(k|P|\lg k|V|)$  process just as in initialization. However, pruning dramatically reduce these calculations because only moves for vertices sharing a edge with the moved vertex will be affected and only those shared edges need to be considered in the computation. Moreover, only those moves involving the move's origin, destination, max or second max parts may be affected and the change in gain can be computed with only those part sizes. Since every vertex will be moved during the pass, total complexity is  $O(k|P||V|\lg k|V|)$ .

Thus far, we have ignored the entropy constraint, yet it does not alter the complexity analysis. From formula 4.2, we see entropy can be computed in  $O(k)$  time. As moves are executed, entropy can be updated in constant time since only sizes of the origin and destination parts need to be known. Finally, recall that we do not execute any moves that violate the entropy constraint. This means we may have to pass over higher gain moves in the tree before finding an allowable move, which is  $O(k|V|)$  to scan through all moves and is less than the  $O(k|P|\lg k|V|)$  process to update gains.

#### 4.3.4 Drawbacks

The FM algorithm can be recursively applied to achieve a k-way partition in  $O(k|P|)$  time, where  $P$  is the set of all pins across all edges (i.e.,  $|P| = \sum_{e \in E} |e|$ , where  $E$  is the set of all edges in the hypergraph). This is a consequence of the fact that the bi-partitioning FM algorithm has complexity  $O(|P|)$  (see Fiduccia and Mattheyses (1982) for analysis). This low complexity is due to two main items. First, it can be observed that a constant number of changes in cut for a single edge is possible during a pass. Second, it makes an assumption that edges will have integer weights. Thus, gains will all be integers and so moves can be indexed by gain in an array of bucket-lists. Under the same assumptions, our k-way algorithm using only the K-1 cut can be implemented with  $O(k|P|)$ . However, in order to support the discount cut, we require  $O(k|P||V|\lg k|V|)$  since the discount cut for an edge can be changed with any of the moves in the pass, leading to the additional  $|V|$  term. We also use a gain tree instead of an array, which leads to the additional  $\lg k|V|$  term. At the same time, use of the tree allows us to support floating point weights, which can be a useful



feature to a user needing more precise control of edge weights. Moreover, use of discounting greatly improves partition quality, as will be shown in section 4.4. Regardless, entropy is a global calculation, so attempting to use recursive bi-partitioning, such as recursively applying FM, would make enforcement of the entropy constraint difficult.

As future work, multi-level k-way partitioning (see section 4.2.1) offers an opportunity to improve run times over our current implementation. While little formal analysis has been performed for multi-level partitioners, experimental evidence has shown dramatic improvements in runtime. Challenges will include determining how to enforce an entropy constraint during the coarsening and uncoarsening phases, as well as properly being able to handle imbalance.

#### 4.4 Experimental Results

Two versions of our algorithm are used. *Hyperpart* only solves for the K-1 objective while *hyperpart-dc* also uses the discount cut. For both, we set  $c_\ell$  and  $c_u$  (from formula 4.2) to be 0.2 below and 0.2 above the desired entropy, respectively. We give *hyperpart* and *hyperpart-dc* this window because at least some tolerance is required to allow movement between parts. This is not unlike the other partitioners. For instance, *hmetis-rb* does not allow a UBFactor less than 1 so that it may have movement around equal-sized parts.

Since none of the comparison methods offer an entropy constraint, we perform multiple runs of each method where we gradually increase their respective imbalance parameters over a range of values. We observe that at high levels of imbalance (low entropy), the comparison partitioners frequently crash or produce partitions with empty parts. So, in some of our tests, some partitioners have no results at low levels of entropy. As mentioned in section 4.2.1, we believe their inability to produce a solution is a result of the difficulty in forecasting the number of vertices needed in each part during the coarsening and uncoarsening phases. This deficiency is also a clear indication that they were designed mainly for balanced partitioning.

For *hmetis-rb*, its UBFactor is varied from 1 to 50 in steps of 0.005. *Hmetis-kway*'s UBFactor is varied from 1 to 200 in steps of 0.015. *Patch*'s *imbal* is varied from 0 to 10

in steps of 0.001. (See section 4.2.2 for UBFactor and imbal definitions.) The maximum level for each of their respective parameters is chosen so that it was beyond the range when they could no longer achieve solutions. The steps are chosen such that the partitioners each produces roughly the same number of solutions and such that lower granularity does not produce meaningful differences in results. For each of these partitioners in each test, the partition selected for comparison is the one with entropy closest to the desired entropy. Clearly, this approach takes much more time than a single run of hyperpart or hyperpart-dc.

#### 4.4.1 Performance with Imbalance

We consider hypergraphs with varying levels of imbalance as shown in Table 4.1. The total number of pins helps quantify partition difficulty as seen in the complexity of FM and other algorithms (see section 4.3.4), so we hold it constant across our test hypergraphs. We generate a single edge for each part in round-robin fashion until we exceed 20000 total pins. For each edge, pins equal to 50% of the part size (with a 2 pin minimum) are generated such that each pin is chosen from the edge’s part with 95% probability and 5% probability of random assignment. These probabilities were determined by experimentation to generate some randomness, but still ensure that the lowest K-1 cut actually follows the specified part sizes, which we call the “true partition.”

For each set of part sizes in Table 4.1, we generate 30 test hypergraphs. Results are shown in Figure 4.4, where we display the average K-1 cut size relative to true partition. We see hmetis-kway has the most limited range, followed by patoh. Hmetis-rb is able to generate solutions at all levels of imbalance, although it sometimes crashes near the higher levels of imbalance in our tests. All partitioners are able to recover the true partition when imbalance is lowest, but as it increases, hyperpart and hyperpart-dc clearly perform better as they return partitions with lower cut size. For part size set #14, we see a spike in the K-1 cut. Observe that #14 has the largest single part size and therefore, the largest edges. In these situations, the discount cut is most beneficial, as we will show in the next section.

Table 4.1: Part Sizes for Test Hypergraphs

#	Part Sizes							Entropy
1	20	20	20	20	20	20	20	2.807
2	17	18	19	20	21	22	23	2.800
3	14	16	18	20	22	24	26	2.778
4	11	14	17	20	23	26	29	2.741
5	8	12	16	20	24	28	32	2.687
6	5	10	15	20	25	30	35	2.610
7	2	8	14	20	26	32	38	2.505
8	2	5	12	20	28	35	38	2.439
9	2	3	6	18	32	39	40	2.298
10	2	2	2	12	32	40	50	2.100
11	2	2	2	2	24	40	68	1.809
12	2	2	2	2	7	28	97	1.401
13	2	2	2	2	2	19	111	1.094
14	2	2	2	2	2	2	128	0.644

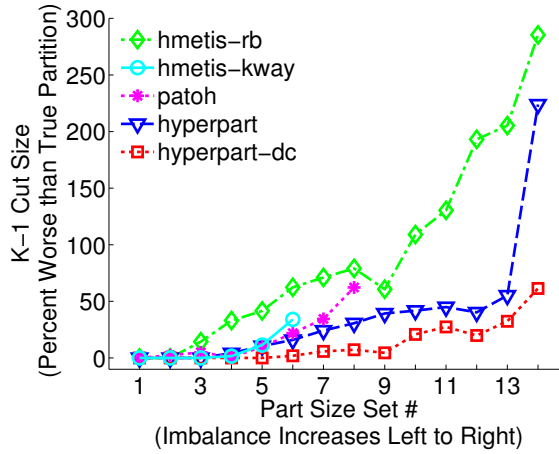


Figure 4.4: Imbalance Test Results. Each tick on the x-axis corresponds to a row in Table 4.1

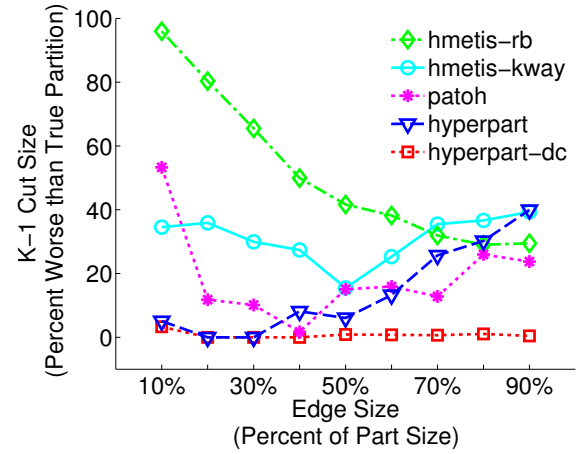


Figure 4.5: Edge Size Test Results

#### 4.4.2 Performance with Edge Size

We fix the part sizes to be #6 in Table 4.1, which is the highest level of imbalance where all partitioners are still able to produce solutions. We use the same hypergraph generation method described in section 4.4.1, but consider graphs with different percentages of pins per edge (instead of just 50%).

Results are shown in Figure 4.5. We see that as edge size grows from 10% to 90% of part size, performance of hyperpart begins to degrade. Meanwhile, hyperpart-dc maintains constant performance, nearly always returning the true partition. As described in section 4.3.2, whereas moves are essentially taken blindly in hyperpart unless they immediately lead to a reduction in K-1 cut, the discount cut guides hyperpart-dc to where the K-1 cut can be reduced.

#### 4.4.3 Real World Datasets

We consider three publicly-available real-world datasets with characteristics described in Table 4.2.

- **Zoo** is from UCI<sup>1</sup> (Bache and Lichman, 2013). Animals are described by one numeric attribute, `num_legs`, and 15 binary attributes, like `produces_milk`, `lives_in_water`, etc. Each value for each attribute is used to form a hyperedge. An expert assignment into 7 groups forms the “actual” partition.
- **Movie** was obtained from Rovi<sup>2</sup>, which backs services like DirecTV and iTunes. We obtain the “significant” movies for 3 genres: Children/Family, Romance and Horror. We form hyperedges using “flags,” which are essentially parental warnings. For example, the family movie *E.T.* has flags “Adult Language,” “Alcohol Consumption,” “Child Classic,” “Scary Moments” and “Watch with your Children.” Some clearly related to family movies, but others could easily apply to a romance or horror.

---

<sup>1</sup><http://archive.ics.uci.edu/ml/>

<sup>2</sup><http://developer.rovicorp.com>

Table 4.2: Real-World Dataset Characteristics

Dataset	Vertices	Parts	Edges	Pins	Max Entropy	Actual Entropy	Actual K-1 Cut
Zoo	101	7	31	1616	2.807	2.391	119
Movie	169	3	32	946	1.585	1.489	22
Track	275	32	527	3412	4.998	4.947	1195

Table 4.3: Real-World Dataset Results

Dataset	Zoo		Movie		Track	
	K-1 Cut	ARI	K-1 Cut	ARI	K-1 Cut	ARI
hmetis-rb	123	0.101	18	0.226	982	0.182
hmetis-kway	×	×	17	0.299	×	×
patoh	101	0.140	14	0.269	851	0.194
hyperpart	101	0.195	18	0.236	831	0.219
hyperpart-dc	100	0.172	13	0.331	833	0.224

- **Track** is the 2011-2012 NCAA division I men’s outdoor track and field schedule, where universities are vertices and track meets are edges connecting them. Conferences, such as the Big Ten, form the “actual” partition. The schedule was obtained from TFRRS<sup>3</sup>, while conferences were obtained from USTFCCCA<sup>4</sup>.

Results are shown in Table 4.3 with × indicating that the partitioner was unable to return a partition near the dataset’s actual entropy. ARI is the Adjusted Rand Index, which measures the agreement between the produced partition and the actual partition on a −1.0 to 1.0 scale, with 1.0 indicating identical partitions. As can be seen by comparing the actual K-1 cut in Table 4.2 with the achieved K-1 cuts in Table 4.3, the partitioners may find cuts that are lower than the actual cut sizes. Most real world datasets are like this because the hypergraph may have a lower cut somewhere other than the actual cut location. For this reason, we do not necessarily expect the partitioners to recover the exact

---

<sup>3</sup><http://tfrs.org>

<sup>4</sup><http://ustfccca.org/infozone>

partition. However, we can expect the output of a good partitioner (with low K-1 cut) to have positive ARI since the clusters in the actual partition should contain more similar objects, and thus have more edges connecting them. So, even though the highest possible ARI of 1.0 will not correspond to the partition with lowest K-1 cut size, we do generally expect lower K-1 cut size to be positively related to ARI.

For all datasets, the best solutions on either the K-1 or ARI measures are returned by hyperpart or hyperpart-dc. In the Track dataset, the edges are relatively small, so hyperpart without the discount cut does as well as (or even slightly better than) hyperpart-dc. However, in the Movie dataset where the edges are large, hyperpart-dc performs better.

#### 4.4.4 Consensus Clustering

Strehl and Ghosh (2003) demonstrate the potential of consensus clustering using a synthetic “8D5K” dataset, where points for 5 clusters are generated in 8 dimensional Euclidean space. We reuse this concept, except whereas clusters were originally all equal size, we test on imbalanced cluster sizes. To generate the dataset, we pick 5 means such that its value along each dimension is chosen uniformly at random from 0.0 to 1.0 but such that no two means are within 0.1 of each other in more than 2 dimensions. The points for each cluster are then generated with values for each dimension again selected randomly, but using a Gaussian pdf with 0.1 standard deviation around the value generated for the mean. Using all 8 dimensions, the points in 8D5K are easily separable. The idea behind the dataset is to use multiple clusterers, each with access to one of the  $\binom{8}{2}$  possible pairs of dimensions, and then combine the clusterings to see if it is possible to recover the true clustering. See Figure 4.6 for an illustration. This process of splitting a dataset in multiple feature subsets, then combining them with a consensus clusterer is known as Feature-Distributed Clustering.

For each of the  $\binom{8}{2}$  pairs of dimensions, we use agglomerative clustering with Ward’s distance metric. Following the HPGA consensus clustering method (as described in section 4.2.3), we form a hypergraph where each of the clusters from each of the 2D clusterers is used to form an edge. Hypergraph partitioning is performed to return the consensus clustering.

Cluster sizes with varying levels of imbalance are shown in Table 4.4. For each setting,

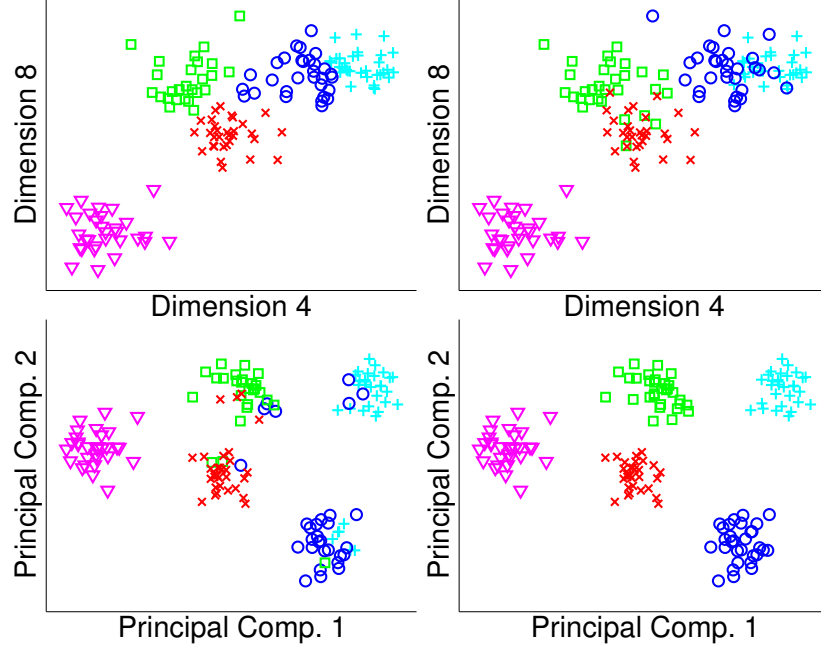


Figure 4.6: 8D5K Dataset. The top figures show data points on 2 of the 8 dataset dimensions. The bottom figures display the same points along the 2 main principal components. Results of a clusterer with access the 2 original dimensions are shown in the left two figures, while the true clustering is shown in the right two figures. Clearly, the overlap in the true clusters along only 2 of the dimensions make it difficult to perform the clustering with complete accuracy, although it is far more accurate than random assignment.

we perform 30 runs, recording the K-1 cut and the minimum, average, and maximum ARIs for the 2-D clusterers. The averages for each over all 30 runs are shown in the table.

Results are shown in Figures 4.7 and 4.8. We see that all partitioners are affected by imbalance, but hyperpart-dc performs best as imbalance increases. Importantly, if we compare the 2-D ARI values from Table 4.4 to results in Figure 4.8, we see that hyperpart-dc has ARI near or above the max ARI of the 2-D clusterers at all levels of imbalance. This indicates that HPGA, using hyperpart-dc as the underlying partitioner, is an effective tool for consensus clustering as quality of the consensus exceeds the inputs, even in imbalanced settings.

Table 4.4: 8D5K Settings

#	Part Sizes					Entropy	Agglom. Clusterer ARI			K-1 Cut
							Min	Average	Max	
1	30	30	30	30	30	2.322	0.423	0.676	0.903	156.9
2	20	25	30	35	40	2.281	0.409	0.673	0.922	153.2
3	10	35	35	35	35	2.220	0.411	0.674	0.910	153.3
4	10	20	30	40	50	2.150	0.402	0.686	0.917	139.5
5	20	20	20	20	70	2.063	0.331	0.632	0.924	144.2
6	8	8	8	63	63	1.728	0.232	0.595	0.906	129.7
7	10	10	10	10	110	1.370	0.164	0.392	0.830	115.7

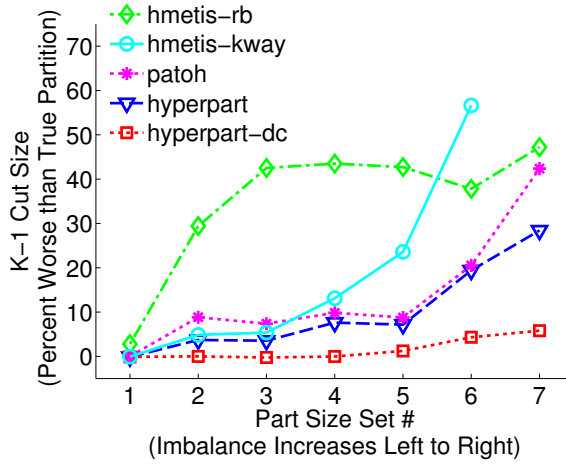


Figure 4.7: Consensus K-1 Cut Results

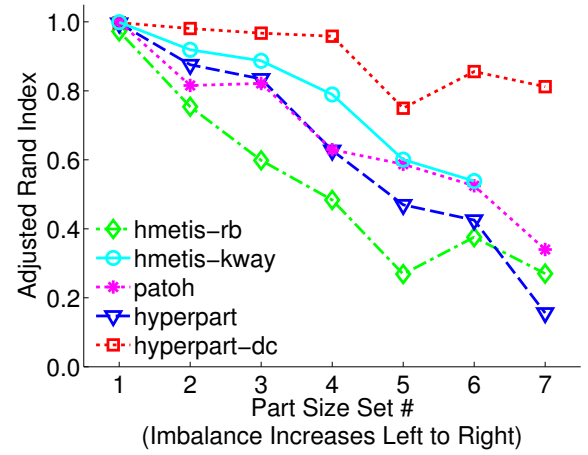


Figure 4.8: Consensus ARI Results



## 4.5 Summary and Future Directions

Many of today’s popular tools and algorithms for hypergraph partitioning have origins in integrated circuit design, where a common specification is that part sizes be roughly equal. Researchers in other domains, particularly consensus clustering, have often considered hypergraph partitioning a poor method when the underlying dataset has imbalanced parts. We suggest that the poor performance is not a result of hypergraph partitioning itself, but rather the constraint definitions used by current tools and algorithms, which essentially measure imbalance only by the largest or smallest part. We argue information-theoretic entropy best measures imbalance and provides an entropy constraint that allows the partitioner find to higher quality solutions for given levels of imbalance. Using a novel “discount” cut heuristic for local optima avoidance along with other known techniques, we devise and implement an algorithm that respects the entropy constraint. Against today’s leading partitioners, we empirically demonstrate our method tends to find lower cut partitions in datasets where the underlying true part sizes are imbalanced. We further show the discount cut is effective in local optima avoidance, particularly in the presence of large edges. Finally, we show that our approach greatly improves the quality of results in consensus clustering when the clusters are imbalanced. These better results are a consequence of the fact that the entropy constraint best allows the partitioner to find “natural” partitions of mixed sizes for any given level of overall imbalance.

As future work, we believe multi-level partitioning with our entropy constraint should be explored in order to improve execution times. Benefits to other consensus clustering algorithms using hypergraph partitioning, such as MCLA, should also be explored. In general, we believe this work broadens the class of problems for which hypergraph partitioning can be considered as an effective method.

## Chapter 5

### Stock Groups

A primary benefit of stock groupings, like those found in industry classifications, is that they provide a simple, high-level view of a universe of stocks. In contrast, pairwise values, such as the ones found in chapter 3, provide very detailed information about similarity, but with such detail, it is difficult to understand larger dynamics, especially for the human user.

However, forming stock groups comes with the cost that some detailed information about individual relationship is lost. For instance, consider TJX Companies (TJX), Ross Stores (ROST) and Abercrombie & Fitch (ANF). All three are almost entirely involved in apparel retail, yet the two discount retailers, TJX and ROST, are intuitively more related than TJX to ANF, a higher-end retailer. In fact, the 2010 correlation of TJX and ROST was 0.703, while TJX and ANF was 0.367. Regardless, a stock grouping may still place them together in a single industry, as GICS does, and these individual relationships are no longer evident.

Still, simplicity can justify the use of stock groupings. An example is the Morningstar Instant X-Ray (Figure 1.1 in chapter 1). The X-Ray allows an investor to quickly see his/her exposures to the different “sectors” of the market. This might help the user to quickly identify if his/her portfolio has concentrated risk into a single area and, if so, to understand how the portfolio should be rebalanced since underweighted sectors are also evident.

At the same time, what defines particular industries is not always well-defined and there are vast possibilities in the formation of these groups. If we consider the 1500 stocks of the S&P 1500, there are  $S(1500,10) = 2.76 \times 10^{1493}$  different ways to form 10 groups, where  $S()$  denotes Stirling numbers of the second kind. Hence, the traditional approach has been to leave the construction of sector/industry taxonomies to experts. Research (e.g., Bhojraj et al. (2003); Chan et al. (2007)) has shown that these experts perform well. Still,

there are indications that improvements could be made. As an example, GICS places Real Estate Investment Trusts (REITs) into the Financials sector, however, these REITs often have characteristics more similar to stocks in other sectors. Consider Marriott International, Inc. (MAR), an S&P 500 constituent that is primarily engaged in operating hotels. In 2010, the S&P 500 company with second highest correlation of daily returns to MAR was Host Hotels & Resorts, Inc. (HST), REIT that owns and operates hotels.<sup>1</sup> In 2010, MAR and HST had returns correlation of 0.818, while the average pairwise correlation of constituents of the S&P 500 was 0.464. Yet, MAR and HST do not appear in the same sector. MAR is classified under Consumer Discretionary, while HST is under Financials because it is a REIT. GICS completely misses the relationship.

This chapter examines using the analyst, correlation and news data to construct stock groups. In section 5.1, two measures for evaluating the quality of groups are developed. In section 5.2, two methods for forming groups are explored: 1) hypergraph partitioning using the algorithm described in chapter 4, and 2) a pipeline approach of a hierarchical clusterer followed by a genetic algorithm applied to the pairwise data of chapter 3. Results are displayed in section 5.3, with comparisons made against three expert-driven classification systems: SIC, FF and GICS.

## 5.1 Evaluation Measures

### 5.1.1 Average Pairwise Correlation

The first quality measure used to compare stock groupings originates in a study by Chan et al. (2007) comparing several industry classification systems. The method evaluates the groups on the basis of stock return co-movement. Two stocks within a group are expected to have higher return correlation than two stocks in different groups. Let  $I$  denote a stock group (i.e., “industry” in the original study). The average pairwise correlation  $\rho_{iI}$  for stock  $i$  in  $I$ , and the average pairwise correlation  $\phi_{iI}$  between stock  $i$  and stocks not in its industry  $I$ , are

---

<sup>1</sup>The S&P 500 company with highest correlation to MAR in 2010 was Starwood Hotels & Resorts Worldwide (HOT) with correlation 0.890.

$$\rho_{iI} = \frac{\sum_{j \in I, j \neq i} d_{ij} \cdot \rho_{ij}}{\sum_{j \in I, j \neq i} d_{ij}} \quad \phi_{iI} = \frac{\sum_{j \notin I} d_{ij} \cdot \rho_{ij}}{\sum_{j \notin I} d_{ij}}$$

where  $\rho_{ij}$  is the Pearson correlation coefficient between returns for stocks  $i$  and  $j$ , and  $d_{ij}$  is the number of days both  $i$  and  $j$  are active. The day-weighting was added by Yaros and Imieliński (2013b) to allow stocks that are delisted to still be included in the computation, thus helping to avoid survivorship bias.

The average intra-industry correlation  $\bar{\rho}_I$  and inter-industry correlation  $\bar{\phi}_I$  for industry  $I$  are:

$$\bar{\rho}_I = \frac{\sum_{i \in I} \rho_{iI}}{|I|} \quad \bar{\phi}_I = \frac{\sum_{i \in I} \phi_{iI}}{|I|}$$

where  $|I|$  is the count of stocks in group  $I$ . Conceptually, if a stock grouping is good,  $\bar{\rho}_I$  will be large and  $\bar{\phi}_I$  will be small.

To aggregate, either a simple average,  $\psi$ , or a weighting by industry size,  $\theta$ , can be used:

$$\psi = \frac{\sum_{I \in \mathbb{I}} (\bar{\rho}_I - \bar{\phi}_I)}{|\mathbb{I}|} \quad \theta = \frac{\sum_{I \in \mathbb{I}} |I| \cdot (\bar{\rho}_I - \bar{\phi}_I)}{\sum_{I \in \mathbb{I}} |I|}$$

where  $\mathbb{I}$  is the set of all industries. The weighted average,  $\theta$ , is generally more preferable since each stock gets equal value.

Unfortunately, optimizing the quality measures  $\psi$  and  $\theta$  directly can easily lead to degenerate solutions. Stocks with the lowest correlation to the market may be placed into their own groups, while all other stocks are placed into a single group. The large group has lower intra-group correlations ( $\rho_I$ ), but the inter-group correlations ( $\phi_I$ ) are much lower, leading to larger  $\psi$  and  $\theta$  measures. Consider Table 5.1. The groups  $\{1,2,3\}$ ,  $\{4,5,6\}$  and  $\{7,8,9\}$  form good natural groups because the stocks within each group are more correlated with each other than with any other stocks. Yet, these natural groups have  $\psi = 0.289$  and  $\theta = 0.289$ , while the degenerate groups  $\{1,2,3,4,5,6,7\}$ ,  $\{8\}$  and  $\{9\}$  have  $\psi = 0.400$  and  $\theta = 0.311$ . These situations easily occur when there are some stocks that have low correlation with the rest of the market. Unfortunately, this happens often in reality as certain companies and entire sectors become distressed and/or experience positive or negative shocks that do not affect the rest of the market.

Table 5.1: Hypothetical Correlations

	1	2	3	4	5	6	7	8	9
1	-	0.6	0.6	0.5	0.5	0.5	0.2	0.0	0.0
2	0.6	-	0.6	0.5	0.5	0.5	0.2	0.0	0.0
3	0.6	0.6	-	0.5	0.5	0.5	0.2	0.0	0.0
4	0.5	0.5	0.5	-	0.6	0.6	0.2	0.0	0.0
5	0.5	0.5	0.5	0.6	-	0.6	0.2	0.0	0.0
6	0.5	0.5	0.5	0.6	0.6	-	0.2	0.0	0.0
7	0.2	0.2	0.2	0.2	0.2	0.2	-	0.3	0.3
8	0.0	0.0	0.0	0.0	0.0	0.0	0.3	-	0.3
9	0.0	0.0	0.0	0.0	0.0	0.0	0.3	0.3	-

The measures  $\psi$  and  $\theta$  have a design that reflects the intentions of statistical cluster analysis, where clusters (i.e., groups) should have high internal homogeneity (as measured by  $\rho_I$ ) and high external separation (as measured by  $\phi_I$ ) (Xu and Wunsch, 2009). However, external separation can clearly be over-emphasized by  $\phi_I$  in the measure. In fact, we propose to simply use  $\rho_I$  alone. Financial practitioners will likely find that the internal similarity of each group is most important to them, rather than focusing on reducing external similarity. Further, there is a strong opportunity cost in placing each stock since  $\rho_I$  will be lower if a stock is placed in an suboptimal group. Therefore, we suggest  $\psi$  and  $\theta$  be replaced by

$$\kappa = \frac{\sum_{I \in \mathbb{I}} \bar{\rho}_I}{|\mathbb{I}|} \quad \gamma = \frac{\sum_{I \in \mathbb{I}} |I| \cdot \bar{\rho}_I}{\sum_{I \in \mathbb{I}} |I|}$$

Further, we prefer the  $\gamma$  measure because each stock gets equal weight, thereby avoiding an imbalancing problem in  $\kappa$  where there is an incentive to form many small groups of highly correlated stocks and put the other stocks into a single large group. We focus on  $\gamma$  for the remainder of this chapter.

### 5.1.2 Average Coefficient of Determination

The second measure used in this dissertation to evaluate group performance is inspired by Bhojraj et al. (2003). As mentioned in section 2.3, Bhojraj et al. originally wished

to evaluate the main contemporary industry taxonomies - SIC, NAICS, Fama-French and GICS. To measure performance, Bhojraj et al. sought to quantify how well movements in individual companies are explained by the contemporaneous average movements of the company's group. This performance can be quantified by the coefficient of determination ( $\mathcal{R}^2$ ) after performing regression

$$R_{i,t} = \alpha_i + \beta_i R_{ind,t} + \epsilon_{i,t}$$

where  $R_{i,t}$  is the return for company  $i$  at time  $t$  and  $R_{ind,t}$  is the equally-weighted average return of all stocks in  $i$ 's industry  $ind$  (i.e., group) at time  $t$ .

Briefly describing the coefficient of determination, it is a number between 0 and 1 intended to describe how well data points fit a model.

$$\mathcal{R}^2 = 1 - \frac{SS_{res}}{SS_{tot}}$$

where  $SS_{res}$  is the sum of the squared residuals from the regression and  $SS_{tot}$  is the sum of squared differences from the mean in the dependent variable (in our case  $R_i$ ).  $\mathcal{R}^2 = 1$  indicates the model explains all variation in the dependent variable, while  $\mathcal{R}^2 = 0$  indicates the model has no explanatory power. For our purposes, the coefficient of determination can be interpreted as an indication of how well the returns of the stocks are explained by the returns of the stock's group using the equation for  $R_{i,t}$  above.

Thus, the average  $\mathcal{R}^2$  values for each grouping can be used for performance evaluation. Higher average  $\mathcal{R}^2$  indicates that the grouping does better at dividing stocks into groups since each stock's movements are better explained by the movements of its group.

## 5.2 Methods

Two approaches to forming groups are presented. The first approach is hypergraph partitioning using analyst data or news data directly to form groups. The second approach uses the pairwise representations from chapter 3 and uses a hierarchical clusterer to form initial groups followed by a genetic algorithm for refinement. The tradeoff between the approaches is that hypergraph partitioning does not have the potential information loss that occurs when looking only at pairs, whereas combining the datasets is more easily accomplished with the pairwise approach (as seen in chapter 3).

### 5.2.1 Hypergraph Partitioning

Analyst coverage can easily be represented as a hypergraph, where companies are represented by vertices and edges represent the analysts that cover those companies. Consider Figure 5.1, which illustrates the process of constructing a hypergraph, then partitioning it. In Figures 5.1a, 5.1b and 5.1c, the analysts from three separate firms are represented. As seen with Cisco and Seagate in Figure 5.1b, it is possible that some companies are not covered by any analysts. This happens frequently in practice. In Figure 5.1d, the complete hypergraph is shown, combining the analyst hyperedges from all firms. In Figure 5.1e, the hypergraph has been partitioned and each part of the partition can be used as a stock group.

Recall from section 2.4 that it can be expected that each analyst will cover a set of highly similar companies. Using this observation, the goal of the hypergraph approach is to minimize the number of occurrences where an analyst’s covered stocks span multiple parts of the partition. That is, we seek a partition that minimizes the edges cut. More specifically, we want to minimize the K-1 cut measure, which is the summation over all edges of the number of parts spanned by an edge minus one (see section 4.2.2 for details). The end result should be a partition of the companies such that the firms in each part are highly similar.

An identical approach can be taken for news articles, where vertices represent companies and each edge represents a single news article, connecting the companies that appear in that article. With the expectation that news articles will tend to cover highly similar companies, a partition minimizing the K-1 cut measure should result in each part having highly similar firms.

To perform partitioning, we use the hyperpart-dc algorithm described in chapter 4 using hmetis to perform the initial partition, which helps reduce run times. We compare the results against three industry classification systems: 1) the US government developed Standard Industry Classification (SIC), 2) the academic Fama-French system (FF), and 3) the commercial Global Industry Classification System (GICS). See section 2.3 for a description of each. To make fair comparisons, we use set the partition size for hyperpart-dc to match

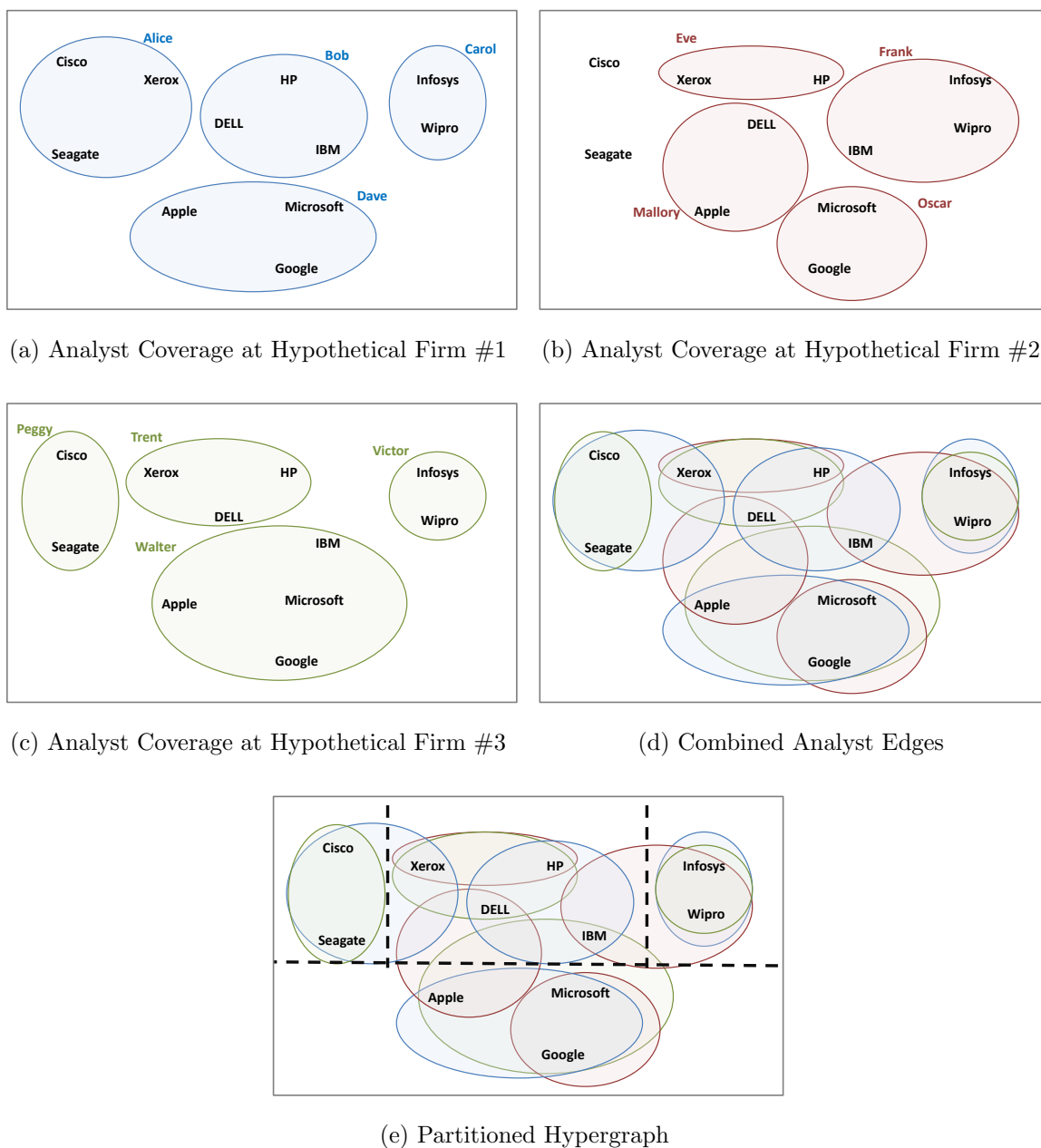


Figure 5.1: Analyst Hypergraph Construction and Partitioning



the number of groups used by the comparison system. For example, we set hyperpart-dc to produce a ten part partition when comparing against GICS sectors because the GICS system has ten sectors. Since each of these comparison systems has multiple levels of granularity (e.g., GICS has sector, industry group, industry and sub-industry), we make comparisons for at each level of granularity. This enables us to assess the hypergraph method’s performance at different levels of aggregation. Additionally, since an unconstrained partitioning leads to degenerate solutions, we target the entropy level of the corresponding comparison system with a tolerance of 0.1 when we perform the experiments.

An additional benefit of the hypergraph approach beyond any improvements in quality is that the approach can perform a partition into any desired number of parts. The expert-driven systems provide only fixed levels of granularity. So, if an investor desired a partition of 100 groups but uses GICS, s/he would either need to accept the 68 industries or 156 sub-industries, or somehow manipulate them to achieve 100 groups.

At the same time, GICS and the other systems do provide additional value in that their groups are labeled. While partitioning the hypergraph might produce a group with Citibank, JPMorgan and Goldman Sachs, it does not label the group. Meanwhile, GICS might label them “financial,” which immediately helps the user understand the business of the constituent companies. While this dissertation does not examine this shortcoming, a variety of approaches could be taken. First, an expert could be used to label the groups, although this runs counter to this dissertation’s goal of greater automation. Second, natural language processing could be used to automatically generate labels. For example, one could use word histograms of company descriptions from their prospectuses to infer major business commonalities within the constituent companies of a group, and apply a label based on high frequency words.

Regardless, hypergraph partitioning (and pairwise conversion methods of section 5.2.2) provide a valuable step towards automation, assuming good performance (which will be seen shortly).

### 5.2.2 Conversion of Pairwise Data into Groups

A disadvantage of the hypergraph representation is that it is not easily extended to combine datasets, particularly correlation. To combine news and analysts, the edges from each dataset could simply be overlaid, but a weighting scheme would likely be necessary since there are many more news articles than analysts. Including correlation is more difficult since it is a pairwise value. Edges of size 2 between each pair of vertices could be added to the graph, with a weight proportionate to the correlation between the two stocks. Yet, this even further complicates weighting the analyst and/or news edges. Moreover, the computation complexity of performing the partition will grow substantially because of the increase in the number of edges.

Given these shortcomings of the hypergraph approach, we also consider forming stock groups from the pairwise values directly. Chapter 3 demonstrated that the analyst and news data are amenable to a pairwise representation, and that combinations of the datasets result in performance improvements. Thus, if groups can readily be formed using pairwise values, then we can expect good performance with the pairwise representation to carry over to the group representation.

At the same time, using pairwise values presents some information loss versus the hypergraph approach. In the pairwise approach, only the strength of similarity between two stocks is quantified, whereas in the hypergraph approach, an edge can represent a bond between any number of stocks. Thus, we may expect some performance degradation if pairwise values are used instead of a pairwise approach. Nevertheless, we still consider this approach because of the greater ease by which the analyst, correlation and news datasets can be combined.

To form groups from pairwise values, an exhaustive search is not feasible because of the enormous number of possible stock groups that can be formed. Fortunately, there are algorithms that can find good approximate solutions relatively quickly. We use a pipeline approach consisting of two algorithms, an agglomerative clusterer and a genetic algorithm.

The agglomerative clusterer begins with each stock in its own group. It then merges the two groups that would lead to the highest overall  $\gamma$  value. It continues this process of

merging until the desired number of groups is reached. This algorithm is relatively fast since the “gain” in  $\gamma$  for merging two groups need only be calculated at initialization and when one of the groups changes (i.e., is merged with another group). Otherwise, no calculation is necessary. The agglomerative clusterer is a greedy algorithm and does not look ahead more than one iteration, so it may have been better to merge different groups in an early iteration to lead to a higher final  $\gamma$ . Such lookahead is computationally infeasible since the algorithm would essentially need to consider a large magnitude of combinations (i.e., the  $S(1500,10) = 2.76 \times 10^{1493}$  described earlier). Instead, we next use a genetic algorithm to refine the solution.

In our genetic algorithm, the sequence of group assignments to stocks is analogous to a DNA sequence of codons. We wish to improve the stock assignments just as a species’ DNA might be improved through natural selection. The algorithm can be described by the standard four stages: 1) Initialization 2) Selection 3) Crossover (a.k.a. Reproduction) and 4) Mutation. Our initialization stage begins by taking the output of the agglomerative clusterer and replicating it into a pool of  $N$  candidate groupings. In the selection phase, we compute  $\gamma$  for each of the  $N$  candidates and select only the top  $P\%$ . In the crossover phase, the selected candidates are placed into parent pools and  $N$  new child candidates are produced by selecting each stock’s group assignment randomly from each of its  $M$  parents. In the mutation phase, these child candidates will have each of their individual stock assignments (i.e., codons) changed with a low probability  $Q\%$ . These candidates are then fed back into the selection phase and the process is repeated. Throughout the algorithm, the best grouping ever seen (i.e., the one with highest  $\gamma$ ) is recorded and if there is no improvement after  $T$  iterations, the algorithm terminates, returning that best grouping. In our experiments, parameter settings are  $N = 1000$ ,  $M = 2$ ,  $P = 25\%$ ,  $Q = 0.5\%$  and  $T = 50$ . Genetic algorithms have been shown to be effective in a variety of settings (see Banzhaf et al. (1998) for an introduction), and we have found ours does well in improving the initial solution produced by the agglomerative clusterer.

The agglomerative and genetic algorithms are applied to the historical pairwise correlations with the hope that the optimal groups of a given year will be predictive of intra-group

correlation for the next year. For analysts and news, we use the same algorithms on their cosine values, using the same quality measure,  $\gamma$ , except replacing pairwise correlation values  $\rho_{ij}$  with the pairwise analyst cosine values or the news cosine values ( $\mathcal{C}_{ij}^A$  or  $\mathcal{C}_{ij}^N$ , respectively, from section 3.1). In a similar fashion, we use the values from the combined analyst, news and historical correlation method of section 3.4.1 to generate another comparative set of groups. To briefly summarize, the steps to compute groups for year  $t$  are:

1. Compute Correlation ( $\rho_{i,j,t-2}$  and  $\rho_{i,j,t-1}$ ), Analyst Cosine ( $\mathcal{C}_{i,j,t-2}^A$  and  $\mathcal{C}_{i,j,t-1}^A$ ), and News Cosine ( $\mathcal{C}_{i,j,t-2}^N$  and  $\mathcal{C}_{i,j,t-1}^N$ ) for previous two years for all pairs of stocks  $i$  and  $j$ .
2. Find parameters  $w_A$ ,  $w_N$ ,  $\varphi_A$  and  $\varphi_N$  that optimize Kendall's tau between last year's correlation ( $\rho_{i,j,t-1}$ ) and the linear combinations of the prior year's ( $t-2$ ) predictors

$$\begin{aligned} \rho_{i,j,t-1} \sim & \rho_{i,j,t-2} + w_A \cdot \left(1 - e^{-\varphi_A \min(a_{i,t-2}, a_{j,t-2})}\right) \cdot \mathcal{C}_{i,j,t-2}^A \\ & + w_N \cdot \left(1 - e^{-\varphi_N \min(m_{i,t-2}, m_{j,t-2})}\right) \cdot \mathcal{C}_{i,j,t-2}^N \end{aligned}$$

3. Using optimized parameters, compute combined similarity value  $s_{i,j,t-1}$  using last year's predictors

$$\begin{aligned} s_{i,j,t-1} = & \rho_{i,j,t-1} + w_A \cdot \left(1 - e^{-\varphi_A \min(a_{i,t-1}, a_{j,t-1})}\right) \cdot \mathcal{C}_{i,j,t-1}^A \\ & + w_N \cdot \left(1 - e^{-\varphi_N \min(m_{i,t-1}, m_{j,t-1})}\right) \cdot \mathcal{C}_{i,j,t-1}^N \end{aligned}$$

4. Run Agglomerative Clusterer on combined similarity values  $s_{i,j,t-1}$  to form candidate stock groups  $\hat{G}$
5. Using  $\hat{G}$  as initialization, run Genetic Algorithm on combined similarity values  $s_{i,j,t-1}$  to form candidate stock groups  $G$
6. Return  $G$  as prediction for year  $t$

### 5.3 Results

In the following pages, we present results of both the hypergraph and the pairwise methods compared against each of the SIC, FF and GICS systems at each level of granularity in their

respective systems. Results are presented for both the intra-group correlation measure ( $\gamma$  in section 5.1.1) and the coefficient of determination measure ( $\mathcal{R}^2$  in section 5.1.2). Because correlations change dramatically from year to year (see Figure 3.12), we do not plot the intra-group correlation values ( $\gamma$ ) directly. Instead, we normalize these values to a range between that year’s average pairwise value between any two stocks and the theoretical maximum  $\gamma$  that could possibly be achieved in that year. We compute this maximum by running the algorithm of section 5.2.2 directly on that year’s correlation data instead of using the previous year’s data. Clearly, this is an approximation of the maximum and is done for visualization purposes only – it is not used in any statistical tests. A zero on the y-axis indicates no improvement over simply choosing groups at random, while a one on the y-axis indicates achievement of the theoretically maximum possible intra-group correlation. For the coefficient of determination  $\mathcal{R}^2$ , raw values are plotted without any normalization.

The figures on the following pages present a great deal of information. To begin, recall from the previous sections that it might be expected that the hypergraph approach will outperform the pairwise approach because of information loss. Consider Figure 5.6, which presents results using the intra-group correlation measure ( $\gamma$ ) at each level of granularity in GICS. At the sector level, which has 10 groups, the hypergraph methods for both analysts and news outperform their respective counterpart pairwise methods. However, as the number of groups increases, the performance of the hypergraph method degrades relative to the pairwise method. At the most granular level, GICS sub-industries, there are approximately 168 groups (the count changes years to year) for the 1500 stocks of the S&P 1500. At this level, the hypergraph methods clearly underperform. We provide two explanations. First, the information loss is less important as the number of groups increases and, correspondingly, the size of the groups decreases. That is, the gap between the size of the group and the size of a pair diminishes, so the information loss is less severe. Second, the difficulty of partitioning the hypergraph increases as the number of groups grows. This is evident in the complexity analysis of hyperpart discussed in section 4.3.3, where the number of groups is a major component of its runtime complexity. This has also been observed with other algorithms, such as k-FM (Sanchis, 1989). So, as the complexity of the partitioning problem increases, the relative quality of the partition produced by the hypergraph approach

degrades. This relationship with the hypergraph approach outperforming the pairwise approach at low levels of granularity but underperforming at high levels of granularity is also evident with the  $\mathcal{R}^2$  measure for GICS in Figure 5.7. With the SIC and FF systems, it also occurs, although FF never really reaches the high number of groups that occur with GICS, so the underperformance of the hypergraph approach is not evident. The highest number of groups in FF is 48<sup>2</sup>.

Evident in all figures is that using only news data underperforms all other methods, regardless of whether hypergraph partitioning or the pairwise method is used on the news data. These results are unsurprising given the results of chapter 3. At the same time, the news data does contain some information. Otherwise, it would have performance comparable to a random set of groups. This is not the case because in all figures displaying intra-group correlation  $\gamma$ , the performance of the news methods are greater than zero.

In contrast to the performance of news, analysts perform much better. Against SIC and FF, the hypergraph method, the pairwise method or both outperform the corresponding SIC or FF groups at each of their respective levels of granularity. Against GICS with the  $\gamma$  and  $\mathcal{R}^2$  measures, analysts underperform at the sector and industry group levels. At the industry level, the analyst pairwise method outperforms GICS on the  $\gamma$  measure with statistical significance ( $p = 0.013$ ) under a paired t-test, although the differences are not significant with the  $\mathcal{R}^2$  measure. At the sub-industry level, the analyst pairwise method outperforms GICS under both the  $\gamma$  measure ( $p = 4.7 \times 10^{-5}$ ) and the  $\mathcal{R}^2$  measure ( $p = 0.014$ ).

Correlation also performs much better than news and performs better than SIC or FF at each of their respective levels of granularity. Against analysts, correlation tends to perform better when there are fewer groups. For instance, at the GICS sector level, correlation outperforms both analyst methods, but at the sub-industry level, correlation underperforms the analyst pairwise method under both the  $\gamma$  measure ( $p = 1.4 \times 10^{-4}$ ) and the  $\mathcal{R}^2$  measure ( $p = 5.1 \times 10^{-5}$ ). Correlation also underperforms the analyst hypergraph method at the GICS sub-industry level, although the differences are significance only with the  $\mathcal{R}^2$  measure

---

<sup>2</sup>There is a 49 group partition available from French's data library (French, 2012), but it was only initiated in 2004.

( $p = 0.004$ ). Against GICS, neither correlation nor GICS does better than the other under the  $\gamma$  measure with statistical significance, except at the sub-industry level where GICS outperforms correlation ( $p = 0.007$ ). With  $\mathcal{R}^2$ , correlation underperforms GICS at all levels of granularity, which may be a reflection of the fact that our hierarchical clusterer and genetic algorithm essentially optimize for  $\gamma$  in the previous year without consideration for  $\mathcal{R}^2$ . In general, correlation tends to perform better than the analyst pairwise method when its performance is weakest – with fewer groups. This matches expectations since, as described in section 3.3.1, analysts tend to cover highly similar stocks and, thus, will be good at creating groups of highly similar companies (i.e., smaller groups), but not be as good at creating larger groups. At the same time, since correlation offers a measure of similarity for all stocks, correlation can do better at forming larger groups.

By combining correlation with analysts and news, the hope is to receive the benefits of each dataset across the spectrum of stock similarity. We wish to combine the performance of analysts on highly similar stocks (plus any potential benefits from news) with the performance of correlation on all other stocks. Indeed, the A+C+N method tends to perform at least as good as the best input method (i.e., analysts or correlation). In the SIC and FF comparisons, the A+C+N method is nearly always the best. Against GICS under the  $\gamma$  measure, neither A+C+N nor GICS performs better than the other at the sector or industry-group level with statistical significance, but A+C+N outperforms GICS at the industry ( $p = 0.001$ ) and sub-industry levels ( $p = 6.9 \times 10^{-4}$ ). Under the  $\mathcal{R}^2$  measure, neither A+C+N nor GICS performs better than the other, except at the sub-industry level ( $p = 0.023$ ).

In general, the pairwise methods tend to do better with the  $\gamma$  measure than with  $\mathcal{R}^2$ . We suggest this occurs because the objective used in the pairwise algorithms more directly matches  $\gamma$  than  $\mathcal{R}^2$ .

The results presented tend to support the results of previous research (especially Bhojraj et al. (2003) and Chan et al. (2007)) comparing the SIC, FF and GICS systems. Although SIC, FF and GICS were not directly compared against each other in our experiments, their performance against the methods presented in this dissertation suggest that GICS has highest quality groups. On the other hand, this dissertation finds that historical correlation

can be used to form groups of higher quality than has been previously implied (especially by Chan et al. (2007)). We believe this is a result of designing an algorithm that directly focuses on optimizing the performance measure  $\gamma$ . Although correlation does not perform as well as GICS, its performance is relatively close. In fact, GICS does not outperform correlation under the  $\gamma$  measure with statistical significance, except at the sub-industry level.

## 5.4 Summary

The focus of this chapter is the formation of stock groups (i.e., a partition of a universe of stocks) with the goal that the stocks in each group be highly related. Two evaluation measures based on previous work are presented. The  $\gamma$  measure essentially focuses on intra-group correlation, while the  $\mathcal{R}^2$  measure focuses on how well the movements of a stock are explained by the averaged movements of the stocks in its group.

Two methods for forming groups are presented. The first is hypergraph partitioning performed using the hyperpart-dc algorithm presented in chapter 4. The hypergraph representation is applied separately to each of the news and analyst datasets, but the vertices represent companies in both cases. For analysts, each edge represents the coverage of an analyst. For news, each edge represents a single news article - uniting the companies co-occurring in the article. The second group formation method is a pipeline approach of a heirarchical clusterer followed by a genetic algorithm. This approach is applied to the pairwise datasets presented in chapter 3: analyst cosine, news cosine, historical correlation, and their combination (A+C+N).

For both analysts and news, the hypergraph approach generally does better than the pairwise approach when forming fewer groups (i.e., group sizes are larger). This is likely a result of less information loss in the hypergraph approach than the pairwise approach. Conversely, the pairwise approach does better when there are more groups (i.e., fewer stocks), which is likely due to a combination of two reasons: 1) the smaller group sizes are closer to the pairs of the pairwise approach, and 2) the computational complexity of hypergraph partitioning is dependent on the number of parts, so more groups means a more



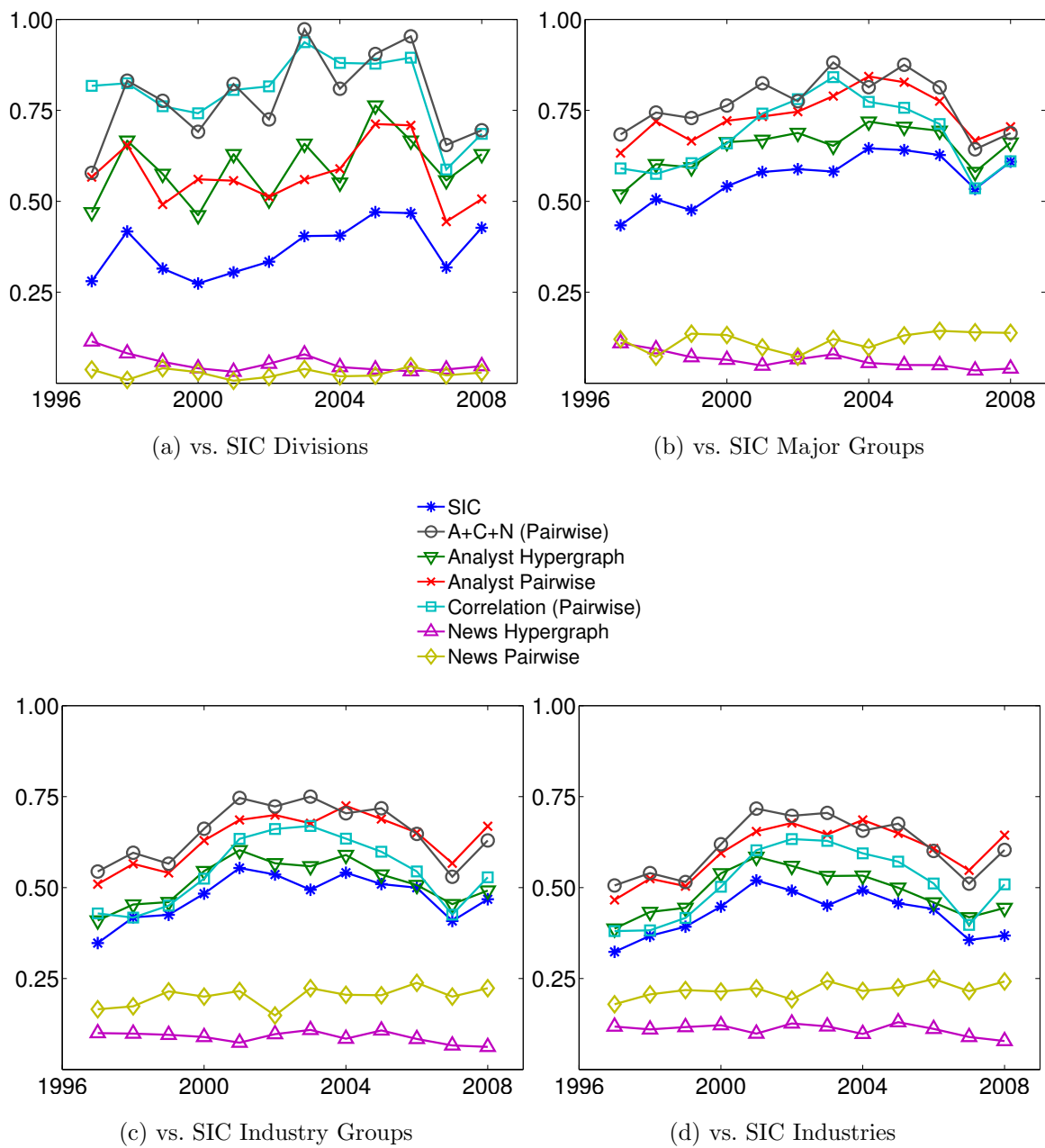


Figure 5.2: Intra-Group Correlation ( $\gamma$ ) compared against SIC

Values on the y-axis range from zero, representing the average intra-group correlation  $\gamma$  for a random partition, to one, representing the theoretical maximum intra-group correlation  $\gamma$  that could have been achieved.

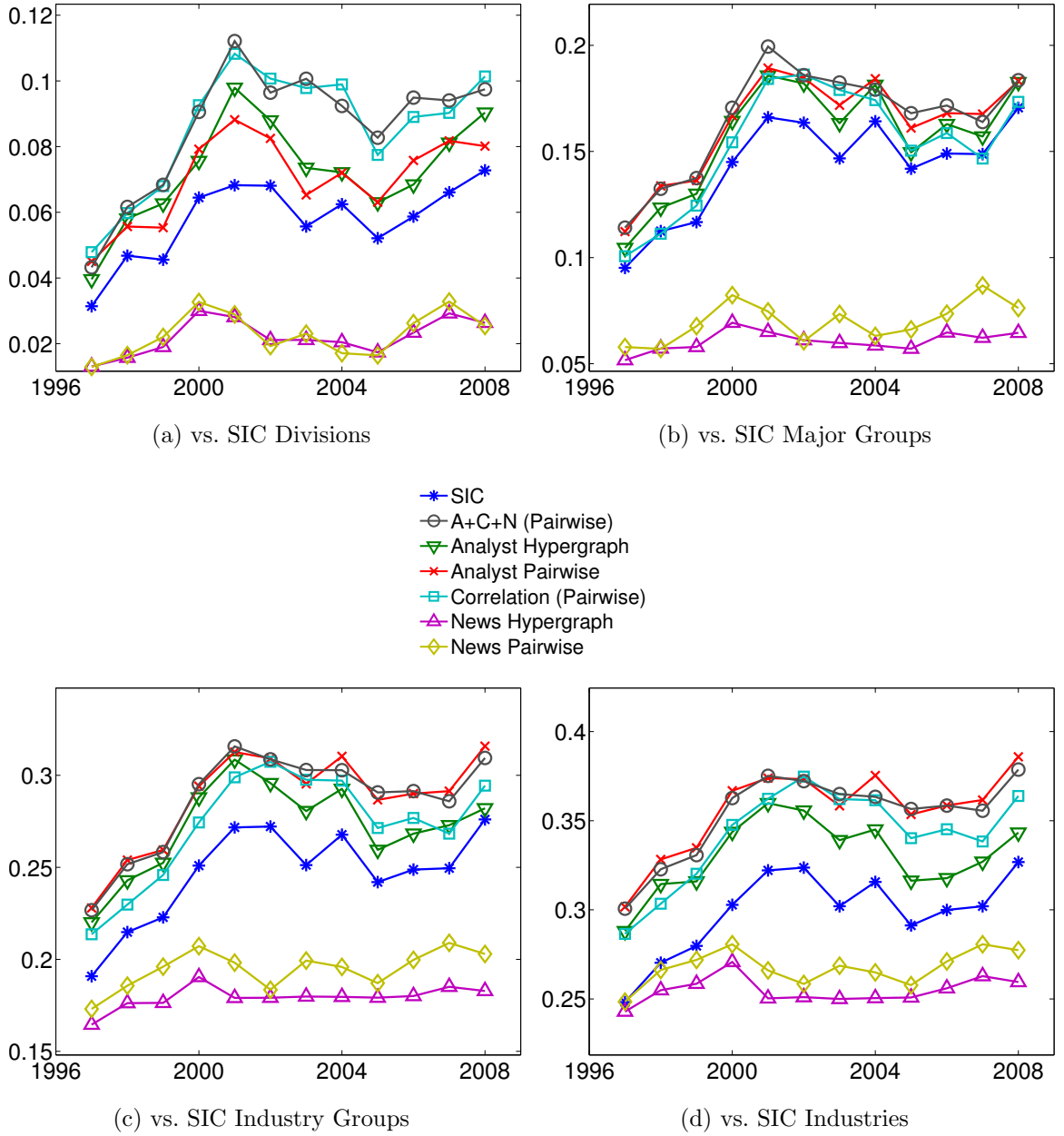


Figure 5.3: Coefficients of Determination ( $\mathcal{R}^2$ ) compared against SIC

Values on the y-axis represent the average  $\mathcal{R}^2$  associated with regressing each stock's daily time series of returns against the averaged returns stock's respective group.  $\mathcal{R}^2$  ranges from zero to one with higher values indicating the stock's returns are better explained by its respective group.

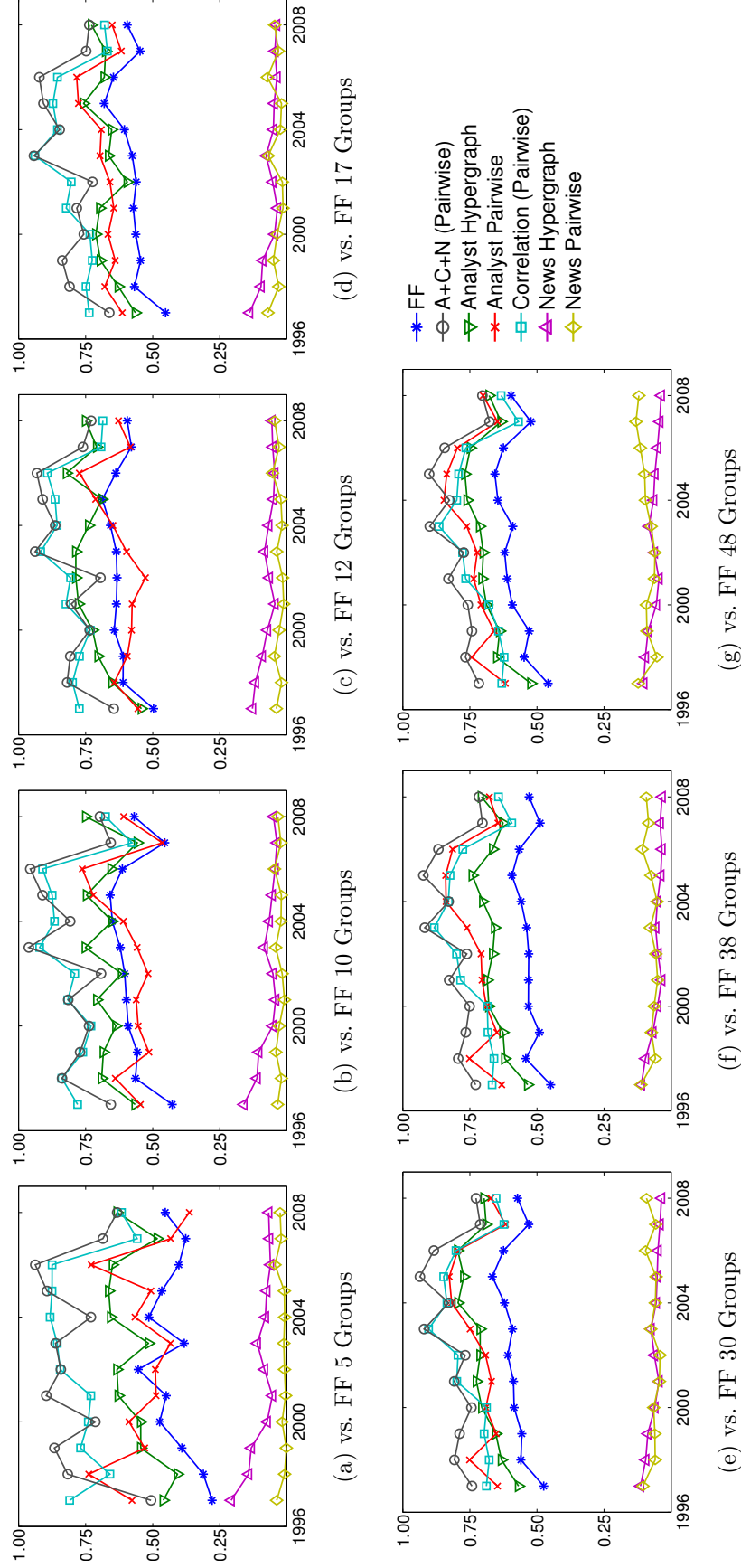


Figure 5.4: Intra-Group Correlation ( $\gamma$ ) compared against Fama-French (FF)

Values on the y-axis range from zero, representing the average intra-group correlation  $\gamma$  for a random partition, to one, representing the theoretical maximum intra-group correlation  $\gamma$  that could have been achieved.

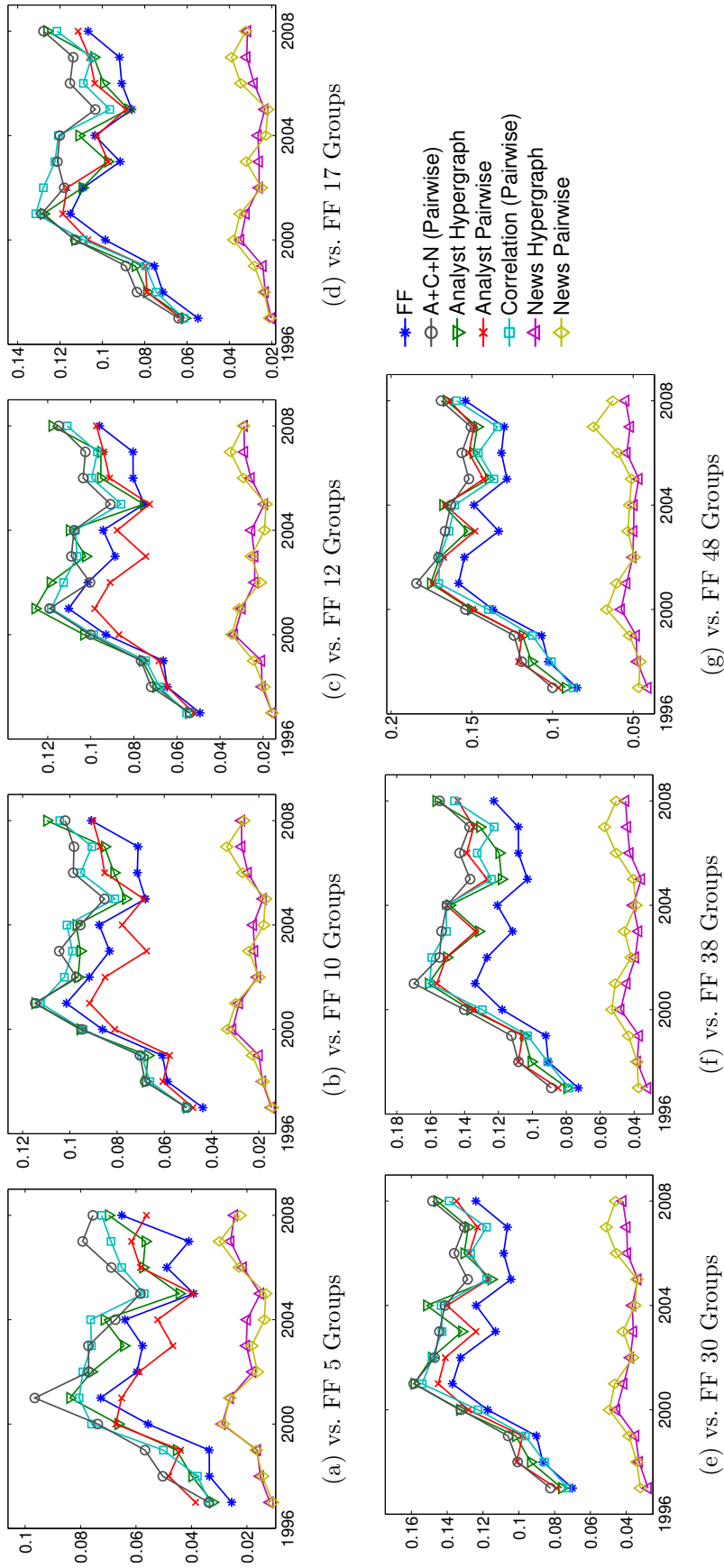


Figure 5.5: Coefficients of Determination ( $\mathcal{R}^2$ ) compared against Fama-French (FF)

Values on the y-axis represent the average  $\mathcal{R}^2$  associated with regressing each stock's daily time series of returns against the averaged returns stock's respective group.  $\mathcal{R}^2$  ranges from zero to one with higher values indicating the stock's returns are better explained by its respective group.

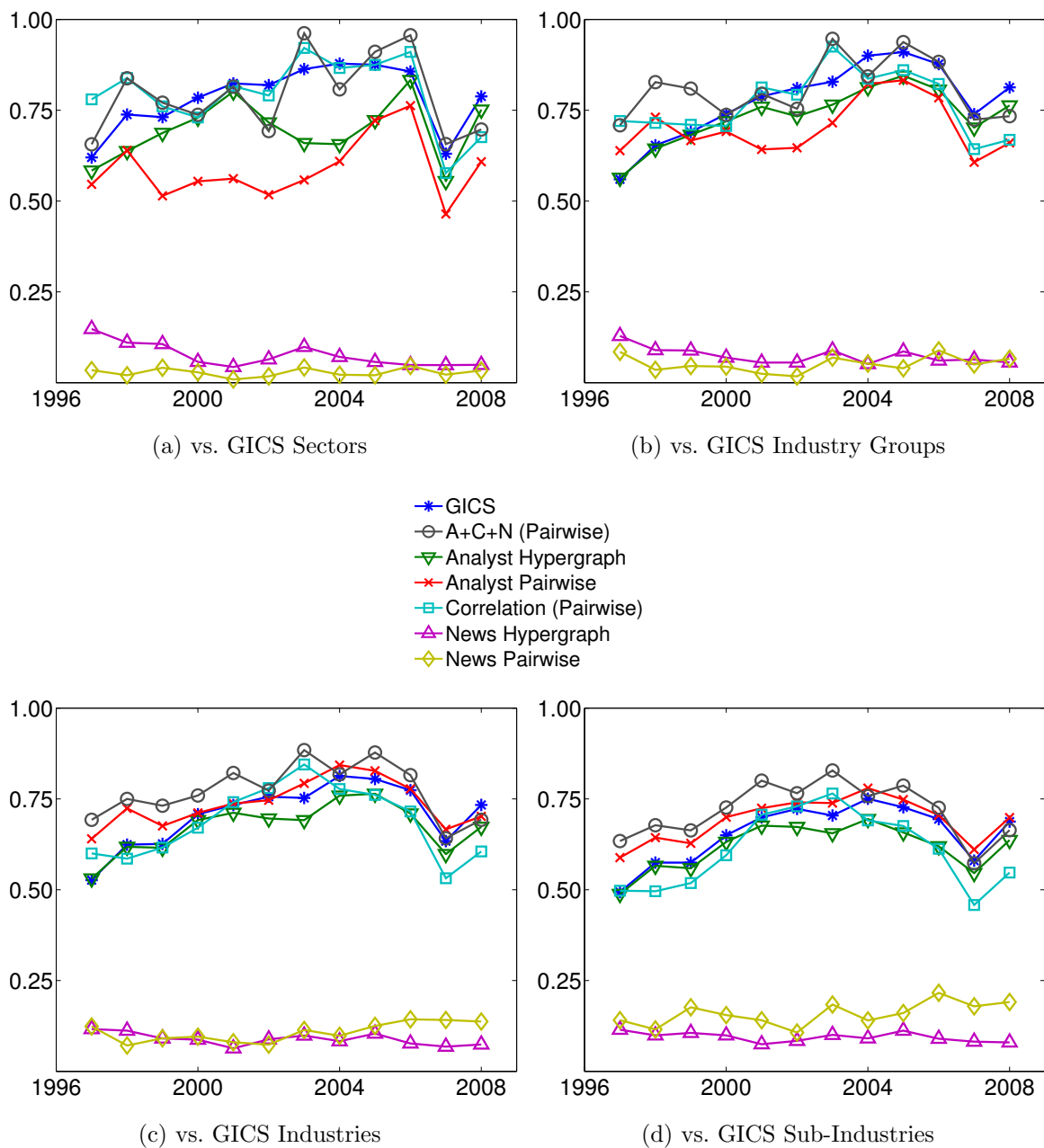


Figure 5.6: Intra-Group Correlation ( $\gamma$ ) compared against GICS

Values on the y-axis range from zero, representing the average intra-group correlation  $\gamma$  for a random partition, to one, representing the theoretical maximum intra-group correlation  $\gamma$  that could have been achieved.

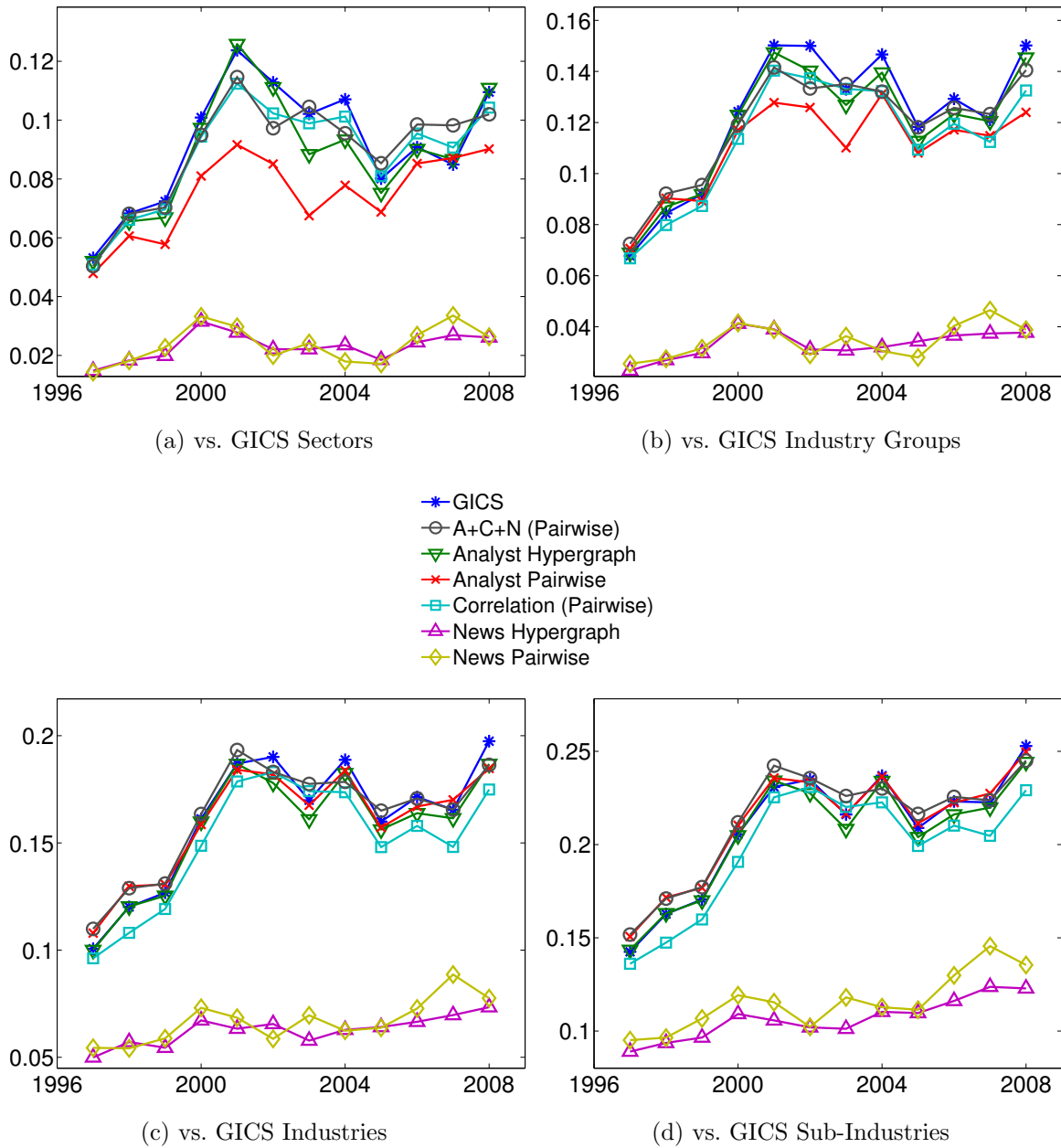


Figure 5.7: Coefficients of Determination ( $\mathcal{R}^2$ ) compared against GICS

Values on the y-axis represent the average  $\mathcal{R}^2$  associated with regressing each stock's daily time series of returns against the averaged returns stock's respective group.  $\mathcal{R}^2$  ranges from zero to one with higher values indicating the stock's returns are better explained by its respective group.

difficult problem for the partitioner.

Among the analyst, correlation and news datasets, news performs worst. Correlation tends to perform better than analysts when forming fewer groups, while analysts tend to do better with more groups. A+C+N does best and generally has performance at least as good as the best of its input datasets (analysts, correlation or news).

These methods are compared against three expert-constructed industry classification systems discussed in section 2.3: SIC, FF and GICS. Comparisons are made at each of their respective levels of granularity. Against SIC and FF, news does worse, but analysts, correlation and A+C+N do better. GICS is the most competitive system, which is in accord with previous research. Against GICS, A+C+N tends to perform at least as well, and often better when more groups are present (i.e., the industry or sub-industry levels). This provides another validation for the methods used to create A+C+N in chapter 3.

While offering performance at least as good as GICS, there are other advantages to the A+C+N approach. First, GICS offers only four levels of granularity while any number of groups can be formed with the A+C+N approach. Second, whereas using GICS carries a reliance on a single expert, the A+C+N approach seeks consensus. The analyst data represents a consensus of brokerage and equity research firms. The news data represents a consensus of news writers. The correlation data represents technical information provided by stock markets, which itself is a consensus of all market participants.

## Chapter 6

### Applications

In chapters 3 and 5, methods were described that used news and analyst data to increase correlation prediction and improve the quality of stock groupings. While these are certainly important to financial professionals, they are not their direct goal. Rather, their most basic objectives are generally to increase returns and to minimize risk. This chapter demonstrates how the improvements made in previous chapters are easily extended to these basic functions. We first consider two areas in depth: section 6.1 describes improvements to diversification, while section 6.2 considers a scenario where an investor wishes to hedge a long position s/he is temporarily unable to unwind. In section 6.3, we present some ideas for further application areas where we have not performed experiments, but believe potential benefits exist.

#### 6.1 Diversification

As described in section 2.1, one of the major advances of the past century in finance was the introduction of Modern Portfolio Theory (MPT) by Markowitz (1952). MPT suggests that risk, as measured by variance of returns, can be reduced by avoiding holding assets with high correlation in the same portfolio. That is, as the pairwise correlations of the individual assets held in the portfolio decrease, the risk of the portfolio should also decrease.

Since future correlations are unknown *a priori*, historical correlation is frequently used as a proxy. In chapter 3, we demonstrated methods to improve prediction of future correlation beyond the simple use of historical correlation by also incorporating analysts and news data. Therefore, it can be expected that these same methods can be used to improve diversification, and thus, reduce risk. This section is devoted to exploring the application of the analyst and news data to the task of diversification.



### 6.1.1 Experimental Setup

The approach taken by this dissertation to achieving diversification is to avoid having pairs of highly similar stocks in the same portfolio. Thus, to quantify the risk of a given portfolio, we use a simple method of counting the number of ordered pairs of stocks in the portfolio that are expected to be highly similar through one (or more) of the pairwise values presented in chapter 3. For example, using the news cosine  $\mathcal{C}^N$ , we count the number of occurrences of an ordered pair  $(c_1, c_2)$  where  $\mathcal{C}^N(c_1, c_2)$  is in the top K highest values for the given stock  $c_1$ . Observe that a set of two stocks  $\{c_a, c_b\}$  may be counted twice - once for  $\mathcal{C}^N(c_a, c_b)$  and once for  $\mathcal{C}^N(c_b, c_a)$ . The expectation is that a higher count of “top K pairs” means the portfolio is less diversified and, thus, more risky. So, these risky portfolios should have higher variances in their returns.

For each year in our dataset, we randomly generate 1,000,000 portfolios of five stocks each from the S&P 1500. Since each stock in the portfolio may itself have varying levels of risk, we weight each stock in the portfolio by the inverse of its historical standard deviation of 20-trading-day returns (approx. one calendar month) over the previous three years. That is, the weight  $w_i$  of stock  $i$  in the portfolio is proportional to  $1/\sigma_{h,i}$ , where  $\sigma_{h,i}$  is the historical standard deviation of stock  $i$ . In the case that there are fewer than twelve 20-trading-day periods of returns in the stock’s history, we give that stock weight corresponding to the average historical standard deviation of the other stocks in the S&P 1500 that year. These weights are normalized such that  $\sum_i w_i = 1$ .

This weighting scheme is intended to balance the portfolio such that each stock contributes equal risk within the portfolio. A true equal-risk-weighted portfolio is known as a “risk parity” portfolio and has recently been an area of active research (see Maillard et al., 2010; Clarke et al., 2013). Analytic solutions are in their nascency, so we use the above approximation (inverse of volatility) as has been done in other research (e.g., Asness et al., 2012). This method would be precise if the portfolio constituents were uncorrelated, but since non-zero correlations exist, we are essentially using the lower bound on each constituent’s risk contribution (Clarke et al., 2013).

Even with our weighting scheme, the portfolios themselves may have varying risk depending on their underlying stocks (even though each is weighted for equal risk within the portfolio). So, to make comparisons between portfolios, we compute the following “risk improvement” measure

$$\Lambda = \frac{\sigma_A}{\sigma_H}$$

where  $\sigma_A$  is the actual standard deviation of 20-trading-day returns of the weighted portfolio over the subsequent 12 periods (essentially monthly returns over the next year). The “homogeneous” standard deviation  $\sigma_H$  is a theoretical value that measures the variance as if the stocks had been perfectly correlated (i.e.,  $\rho_{ij} = 1$  for all pairs of stocks  $i, j$ )

$$\sigma_H^2 = \sum_{i=1}^n \sum_{j=1}^n w_i w_j \sigma_i \sigma_j \rho_{ij} = \sum_{i=1}^n \sum_{j=1}^n w_i w_j \sigma_i \sigma_j$$

Of course, most stocks are not perfectly correlated, so  $\Lambda$  will range between zero and one with lower values indicating less risk. A value of 0.5 means that the risk (as measured by variance) is half the risk that would be encountered if the stocks were perfectly correlated.

For comparison, we also consider a common method of achieving diversification: using an industry classification. In this approach, risk can be quantified as the number of GICS sectors a portfolio spans. More sectors means lower risk. For a five stock portfolio, the most diversified portfolio will have each stock from a different sector, so five sectors are spanned. The most risky portfolio will have all five stocks from a single sector. Such an approach is frequently espoused as a basic approach to investing (e.g. Cramer (2005)). In fact, Nasdaq’s financial glossary defines “sector diversification” as “constituting of a portfolio of stocks of companies in each major industry group.”<sup>1</sup>

In practice, other investment dimensions should be considered for diversification. For example, risk can be lowered by investing across multiple asset classes (e.g., bonds, commodities, real estate, etc.) and across countries. However, this does not diminish the importance of these experiments. Diversification should still be applied within the single-country equity portion of a larger investment portfolio (U.S. Securities and Exchange Commission, 2009).

---

<sup>1</sup><http://www.nasdaq.com/investing/glossary/s/sector-diversification>

### 6.1.2 Results

We first consider the validity of quantifying portfolio risk by counting the number of top  $K$  pairs. Figure 6.1 depicts results for year 2000 (other years are similar) using  $K = 50$ . As can be seen in the figure, an increase in the count of instances where a stock is in the top  $K$  of another stock generally leads to an increase in the risk of the portfolio (as measured by  $\Lambda$ ). This relationship is present for all predictors: analyst cosine (Analyst), historical correlation (Correlation), news cosine (News) and the combination method (A+C+N) described in section 3.4.1.

Since investors will typically seek a portfolio with as little risk as possible (given equal expected return), we consider the difference between portfolios with no top  $K$  pairs versus portfolios with one or more top  $K$  pair. Results are shown in Table 6.1. In all but two cases, having zero top  $K$  pairs leads to lower risk.

Thus, the notion of counting top  $K$  pairs to quantify risk has merit. We also consider the traditional approach using sectors. In Figure 6.2, it is evident that the fewer sectors the portfolio spans, the higher the risk. This matches traditional expectations that a well-diversified portfolio will span multiple sectors.

### Portfolio Selectivity

Before the risk results of the top  $K$  and traditional sector approaches can be compared, one must consider their selectivity. If one approach returns fewer portfolios than the other, it may be considered an unfair comparison since it can be easier to reduce risk by simply being highly selective. Figure 6.3 displays, for each possible count of sectors, the percentage of the 1,000,000 randomly generated portfolios that have the given count. The riskiest portfolios (one sector) are rare. On the other hand, the most diversified portfolios (five sectors) comprise approximately 20%. Most portfolios lie between these extremes and span two, three or four sectors.

Using the top  $K$  approach, the selectivity is largely controlled by  $K$  – higher  $K$  means higher selectivity and, thus, fewer portfolios. Consider Figure 6.4, which displays the relationship between the value of  $K$  and the percentage of portfolios with zero top  $K$  instances.

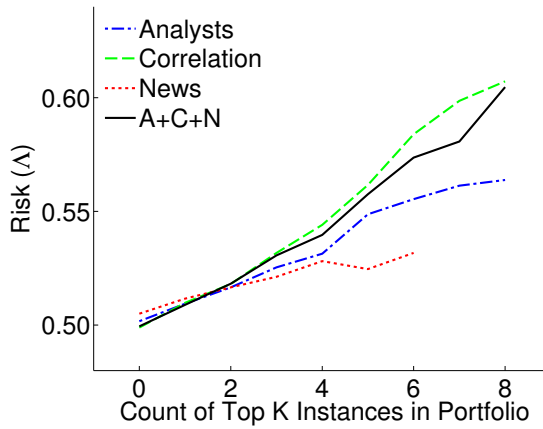


Figure 6.1: Relationships between Top K Counts and Risk

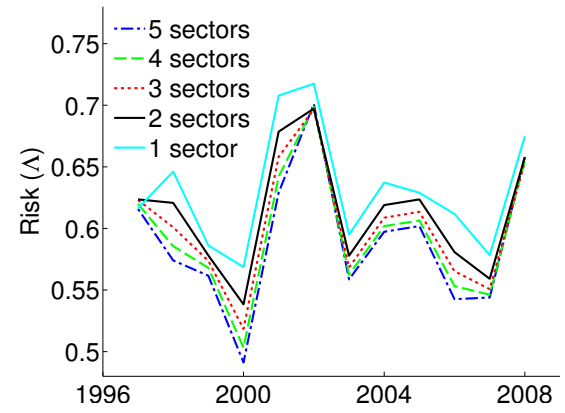


Figure 6.2: Risk by GICS Sector Count

Table 6.1: Risk by Top K Counts

Year	Analysts		Correlation		News		A+C+N	
	zero	one+	zero	one+	zero	one+	zero	one+
1997	<b>0.6178</b>	0.6252	<b>0.6144</b>	0.6278	<b>0.6185</b>	0.6303	<b>0.6157</b>	0.6265
1998	<b>0.5870</b>	0.5917	<b>0.5853</b>	0.5923	<b>0.5876</b>	0.5935	<b>0.5855</b>	0.5924
1999	<b>0.5657</b>	0.5754	<b>0.5641</b>	0.5746	0.5686	<b>0.5679</b>	<b>0.5640</b>	0.5749
2000	<b>0.5017</b>	0.5169	<b>0.4989</b>	0.5162	<b>0.5050</b>	0.5156	<b>0.4994</b>	0.5164
2001	<b>0.6439</b>	0.6495	<b>0.6417</b>	0.6505	<b>0.6449</b>	0.6501	<b>0.6418</b>	0.6507
2002	<b>0.6955</b>	0.7077	<b>0.6945</b>	0.7041	<b>0.6980</b>	0.7031	<b>0.6947</b>	0.7041
2003	<b>0.5613</b>	0.5721	<b>0.5596</b>	0.5696	<b>0.5630</b>	0.5728	<b>0.5597</b>	0.5697
2004	<b>0.6006</b>	0.6114	<b>0.5984</b>	0.6103	<b>0.6015</b>	0.6204	<b>0.5984</b>	0.6105
2005	<b>0.6058</b>	0.6148	<b>0.6006</b>	0.6186	<b>0.6080</b>	0.6093	<b>0.6012</b>	0.6181
2006	<b>0.5539</b>	0.5601	<b>0.5506</b>	0.5623	0.5560	<b>0.5509</b>	<b>0.5506</b>	0.5625
2007	<b>0.5451</b>	0.5542	<b>0.5473</b>	0.5478	<b>0.5451</b>	0.5696	<b>0.5470</b>	0.5482
2008	<b>0.6523</b>	0.6644	<b>0.6546</b>	0.6567	<b>0.6541</b>	0.6691	<b>0.6543</b>	0.6573

K=50. 'zero' columns indicate no top K instances present in the portfolio. 'one+' indicates one or more top K instances are present.

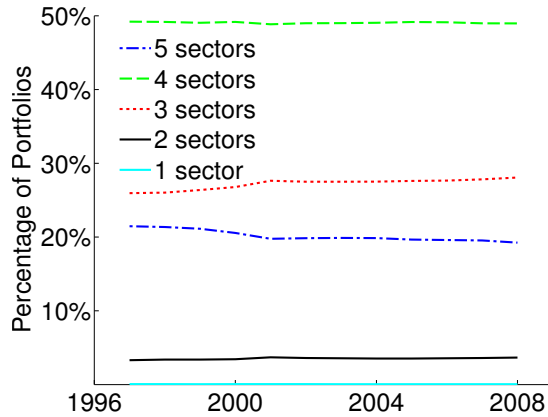


Figure 6.3: GICS Selectivity

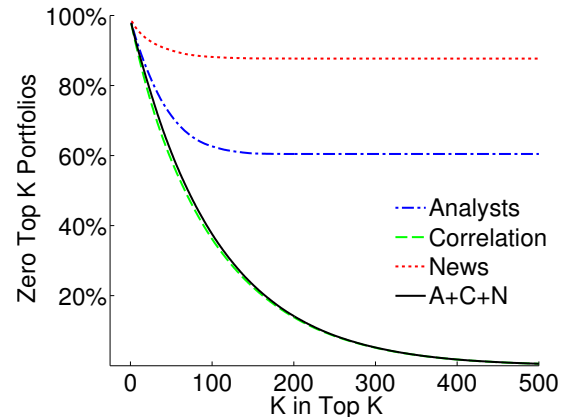


Figure 6.4: Top K Selectivity (year 2000 shown)

As expected, increasing  $K$  decreases the number of portfolios. The figure also demonstrates that the analyst and news portfolios have relatively high asymptotic minima. These minima occur because of the nature of the analyst and news datasets as described in section 3.3.1. For most pairs of stocks, there are no analysts covering both. Likewise, most pairs of stocks do not co-occur in any news stories. Thus, the analyst and news data can be used to eliminate the portfolios containing the pairs of highly similar stocks, but this has limits as  $K$  is increased and portfolios containing moderately related stocks would be next to be eliminated. The analyst and news data have little information in this regard. In fact, they never reach the 20% selectivity of five-sector GICS portfolios, so comparisons are not appropriate. However, since correlation can be computed for any pair of stocks (assuming a historical price time series exists), it can achieve significant levels of selectivity. This carries over to the combination method ( $A+C+N$ ), since correlation is one of its components.

### Comparison with GICS

To compare GICS with correlation, for each year we use the value of  $K$  for correlation that produces the number of portfolios with zero top  $K$  pairs that is closest to the number of GICS portfolios with stocks from 5 sectors. The same approach is applied to compare with  $A+C+N$ . Results are shown in Table 6.2. As seen in the table, none of the methods clearly outperform the other methods. Under paired  $t$ -tests, none are significantly different than

the others. This indicates the correlation and A+C+N approaches have performance that is comparable to GICS.

### **Combination with GICS**

As with the analysts, correlation and news datasets, a natural question is whether GICS can be combined with the other datasets to make improvements. To consider this question, we examine portfolios that each would suggest are most highly diversified. That is, we examine the portfolios that are in the intersection of the set of portfolios with five GICS sectors and the set of portfolios that have zero Top K pairs. Inevitably, this increases selectivity and means that the set of satisfying portfolios will be smaller. However, an investor may want to know which portfolios satisfy all approaches since these are likely the are most diversified.

Results are shown in Table 6.3 using  $K=50$ . In nearly every year (except 2002 and 2008), the risk is reduced by including GICS. The improvements for each are significant under a paired t-test ( $p = 0.010$ ,  $p = 0.012$ ,  $p = 0.001$  and  $p = 0.014$  for analysts, correlation, news, and A+C+N, respectively). Additionally, the combinations also generally improve results over just using GICS alone. Results are significant under paired t-tests comparing GICS to each combination with GICS ( $p = 1.9 \times 10^{-4}$ ,  $p = 2.7 \times 10^{-4}$ ,  $p = 0.014$  and  $p = 3.8 \times 10^{-4}$  for analysts, correlation, news, and A+C+N, respectively).

## **6.2 Long Position Hedging**

This section considers a scenario where an investor holds a long position in a single stock and fears the stock may lose significant value due to a market or sector decline, but it is unable to sell his/her position in that stock for a period of time. This inability to trade could be a result of owning shares that have not yet vested as part of a compensation package. Alternatively, the investor may be prohibited from trading because s/he is an employee with material information or is an executive. In such situations, the value of the stock position may be a significant part of the investor's wealth, so proper hedging is critical to preserving value. Moreover, these investors might easily be barred from trading in the stock's derivatives, such as put options, so the most straightforward hedging possibilities

Table 6.2: Diversification Comparison against GICS

Year	GICS	Correl	A+C+N
1997	0.6158	<b>0.6094</b>	0.6098
1998	<b>0.5740</b>	0.5830	0.5827
1999	0.5616	0.5609	<b>0.5608</b>
2000	<b>0.4912</b>	0.4926	0.4921
2001	<b>0.6290</b>	0.6375	0.6374
2002	0.7011	<b>0.6902</b>	0.6902
2003	0.5586	0.5556	<b>0.5556</b>
2004	0.5973	0.5931	<b>0.5929</b>
2005	0.6019	<b>0.5906</b>	0.5912
2006	<b>0.5425</b>	0.5449	0.5450
2007	<b>0.5439</b>	0.5472	0.5470
2008	0.6579	0.6524	<b>0.6519</b>

Table 6.3: Combinations with GICS

Year	GICS	Analysts		Correlation		News		A+C+N	
		with	w/o	with	w/o	with	w/o	with	w/o
1997	0.6158	<b>0.6158</b>	0.6178	<b>0.6125</b>	0.6144	<b>0.6147</b>	0.6185	<b>0.6143</b>	0.6157
1998	0.5740	<b>0.5733</b>	0.5870	<b>0.5716</b>	0.5853	<b>0.5736</b>	0.5876	<b>0.5721</b>	0.5855
1999	0.5616	<b>0.5610</b>	0.5657	<b>0.5604</b>	0.5641	<b>0.5619</b>	0.5686	<b>0.5605</b>	0.5640
2000	0.4912	<b>0.4906</b>	0.5017	<b>0.4876</b>	0.4989	<b>0.4905</b>	0.5050	<b>0.4890</b>	0.4994
2001	0.6290	<b>0.6289</b>	0.6439	<b>0.6273</b>	0.6417	<b>0.6291</b>	0.6449	<b>0.6275</b>	0.6418
2002	0.7011	0.7002	<b>0.6955</b>	0.6999	<b>0.6945</b>	0.7006	<b>0.6980</b>	0.7002	<b>0.6947</b>
2003	0.5586	<b>0.5581</b>	0.5613	<b>0.5571</b>	0.5596	<b>0.5579</b>	0.5630	<b>0.5572</b>	0.5597
2004	0.5973	<b>0.5969</b>	0.6006	<b>0.5945</b>	0.5984	<b>0.5957</b>	0.6015	<b>0.5947</b>	0.5984
2005	0.6019	<b>0.6015</b>	0.6058	<b>0.5996</b>	0.6006	<b>0.6021</b>	0.6080	<b>0.6000</b>	0.6012
2006	0.5425	<b>0.5426</b>	0.5539	<b>0.5404</b>	0.5506	<b>0.5432</b>	0.5560	<b>0.5404</b>	0.5506
2007	0.5439	<b>0.5436</b>	0.5451	<b>0.5454</b>	0.5473	<b>0.5421</b>	0.5451	<b>0.5454</b>	0.5470
2008	0.6579	0.6568	<b>0.6523</b>	0.6566	<b>0.6546</b>	0.6569	<b>0.6541</b>	0.6569	<b>0.6543</b>

are eliminated.

To hedge the investor’s position, we use short sales of similar stocks. An ideal hedge portfolio would have future time series of returns such that every movement in the long stock is offset by an identical movement in the hedge portfolio. In practice, such a portfolio is unachievable because each stock’s returns will be affected by its own specific events. Nevertheless, one can seek to obtain a hedge portfolio as close to the ideal as possible by using the stocks most similar to the long stock. To select these stocks, we consider using the pairwise values described in chapter 3: analyst and news cosine values, correlation and their combination (A+C+N). We compare these hedging methods against each other and against the baselines of hedging with a market index ETF (SPY) and using a sector/industry scheme, GICS, to select a portfolio of stocks.

### 6.2.1 Experimental Setup

To simulate hedging for a variety of time periods and stocks, we generate 100,000 runs, where each run randomly selects a start date from the time period of 1997 to 2008 and a single stock from the S&P 1500 constituents on that start date. We assume the investor wants to hedge a long position in that stock over the next 125 trading days (approx. six months). To hedge, we consider several portfolios. First, we use SPY, an ETF tracking the S&P 500.<sup>2</sup> Second, we construct a portfolio using the GICS taxonomy. Ten stocks are randomly selected from the long stock’s sub-industry. If there are fewer than ten stocks, we use stocks from the long stock’s industry. If there are still too few, we use the industry group and, finally, the sector. Third, we consider stocks selected by the similarity matrices, as computed through correlation, analyst cosine values, news cosine values, or the optimal combination thereof (as described in section 3.4.1). We select ten stocks with largest similarity values to the investor’s long stock. We did not perform analysis to determine if ten is an optimal number of stocks, but such an optimization would anyway be tangential to our main task, which is to evaluate our similarity values as a means to select stocks for hedging.

For each run and each hedging method, we perform the following regression over the

---

<sup>2</sup>An ETF tracking the complete S&P 1500 was not used because such an ETF did not exist until 2004 when ITOT, the iShares Core S&P Total U.S. Stock Market ETF (formally called the The iShares S&P 1500 Index Fund) was introduced. Additionally, ITOT has much less daily volume than SPY.



500 trading days (roughly two calendar years) prior to the run's start date:

$$r_{s,t} = \alpha + \beta \cdot r_{h,t} + \epsilon_t \quad (6.1)$$

where  $r_{s,t}$  and  $r_{h,t}$  are 20-trading-day returns for the stock and the hedge portfolio at time  $t$ , respectively, and  $\epsilon_t$  is the error at time  $t$ . The interval of 20 trading days is roughly one calendar month and means that the regression is over 25 points. For every dollar of the long stock position, we short  $\beta$  dollars of the hedge portfolio (i.e., we hedge in a beta-neutral fashion).

For simplicity, we ignore trading costs like commissions and market impact. For short sales, we assume collateral of 100% of the value of the short is due at each trade's inception. Borrow costs, interest on collateral and all other fees or income are not considered. We do not rebalance the hedge portfolio, even in the case that a hedge stock is delisted. In the case that the long stock is delisted, we close the hedge portfolio on the delist date and use the associated returns in our results.

We use a simplistic scenario for two main reasons. First, accounting for many of the real-world effects that would impact the hedging scenario requires additional data that can be difficult to obtain (e.g., historical borrow rates for individual stocks). Second, accounting for these effects can require using complex models. These models are an active area of research (especially for market impact costs) and choosing a specific model easily invites criticism that may taint our results in their entirety. Since our intention is to portray the potential benefits of our methods in similar stock selection, we stick to a simple scenario such that stock selection should be the primary driver of any witnessed benefits. Nevertheless, it is important to realize that results may be significantly different in practice. In particular, shorting stocks can incur significant costs, especially for the stocks of distressed companies. The environment for borrowing shares to sell stocks short is itself a market and if a company is in distress, the demand to borrow the stock can outstrip supply, leading to high fees or even situations where the stock is impossible to borrow. These factors should be considered before hedging in practice.

### 6.2.2 Results

Table 6.4 contains results over the 100,000 runs for the entire 1997-2010 time period. Table 6.5 focuses on the Global Financial Crisis and has the subset of runs with an initiation between January 2007 and August 2008 (recall that each run is 125 trading days in duration, so trades may end as late as early March 2009). The following items are displayed:

**AvgRet** The arithmetic mean of returns.

**StDev** The sample standard deviation of returns.

**MAD** Mean average deviation from the average return.

**DDown** The largest drawdown: the minimum return (i.e., the worst case)

**VaR[5]** Value at Risk at the 5% threshold: the fifth percentile value when returns are sorted least to greatest.

**ES[5]** Expected Shortfall at the 5% level: the average of the lowest five percent of returns.

**Roy[X]** Roy's Safety-First Criterion (introduced by Roy (1952)): the probability that returns are less than X%. (-25%, -50%, -75% & -100% are shown.)

As seen in Table 6.4, use of hedging reduces risk in multiple measures, including StDev, MAD, VaR[5], ES[5], Roy[-25] and Roy[-50]. However, it is evident in the largest drawdown (DDown) and in the most extreme threshold for Roy's safety-first criterion Roy[-100], that the use of shorts for hedging actually increases "extreme event" risk. Losses to the long position are bounded at 100% if the stock price falls to zero, which is why the unhedged strategy's biggest drawdown is close to -100%. At the same time, losses to the short portfolio are theoretically unbounded since the price of the portfolio can continually rise. This risk can be mitigated (though not eliminated) through stop loss orders on the shorted stocks.

When viewing the average return (AvgRet) in Table 6.4, the reduction in risk may not seem worth the reduction in return. Stock values tend to rise in general (Siegel, 2008) and, thus, short positions should be avoided. However, if one has a strong belief that a market or sector downturn is imminent, use of a short portfolio can reduce risk and preserve value as seen with the crisis period in Table 6.5. Both the risk factors are reduced and the average return (AvgRet) is higher with hedging. The difference is further evident in the histograms of Figures 6.5 and 6.6, where the variance (i.e., risk) is lower in both periods for hedging

Table 6.4: Top 25% S&amp;P 1500 – Hedge Results During Entire 1997-2010 Period

	Unhedged	SPY	GICS	Analyst	Correl	News	A+C+N
AvgRet	<b>3.96%</b>	1.72%	0.40%	0.43%	1.03%	0.36%	1.04%
StDev	31.35%	28.57%	25.78%	24.91%	25.55%	26.99%	<b>24.76%</b>
MAD	21.39%	19.35%	17.00%	16.24%	16.67%	18.16%	<b>16.11%</b>
DDown	<b>-99.89%</b>	-111.87%	-196.60%	-185.73%	-152.95%	-171.89%	-161.59%
VaR[5]	-44.16%	-39.30%	-36.18%	-34.43%	-34.74%	-38.81%	<b>-33.41%</b>
ES[5]	-60.16%	-53.74%	-52.53%	-50.26%	-50.49%	-55.12%	<b>-48.72%</b>
Roy[-25]	12.97%	12.54%	10.54%	9.41%	9.69%	12.10%	<b>8.98%</b>
Roy[-50]	3.62%	2.44%	2.08%	1.75%	1.86%	2.54%	<b>1.64%</b>
Roy[-75]	0.74%	0.49%	0.51%	0.43%	0.46%	0.56%	<b>0.35%</b>
Roy[-100]	<b>0.00%</b>	0.02%	0.12%	0.10%	0.08%	0.09%	0.07%

Table 6.5: Top 25% S&amp;P 1500 – Results During 2007 - 2008 Crisis Period

	Unhedged	SPY	GICS	Analyst	Correl	News	A+C+N
AvgRet	-12.21%	-0.99%	-3.19%	-1.98%	0.54%	-2.29%	<b>0.95%</b>
StDev	28.07%	24.14%	21.35%	<b>20.41%</b>	21.55%	22.61%	20.88%
MAD	21.64%	17.66%	15.64%	<b>14.95%</b>	15.48%	16.57%	15.15%
DDown	-99.89%	-99.72%	<b>-98.44%</b>	-101.54%	-131.84%	-103.53%	-127.52%
VaR[5]	-62.84%	-39.49%	-38.09%	-34.91%	-33.58%	-38.53%	<b>-32.10%</b>
ES[5]	-75.11%	-53.99%	-51.95%	-48.17%	-48.86%	-51.69%	<b>-45.99%</b>
Roy[-25]	28.90%	13.41%	12.98%	10.60%	9.20%	13.29%	<b>8.26%</b>
Roy[-50]	10.09%	2.40%	2.04%	1.49%	1.67%	1.97%	<b>1.20%</b>
Roy[-75]	1.99%	0.55%	0.41%	0.33%	0.35%	0.37%	<b>0.28%</b>
Roy[-100]	<b>0.00%</b>	<b>0.00%</b>	<b>0.00%</b>	0.04%	0.08%	0.04%	0.06%

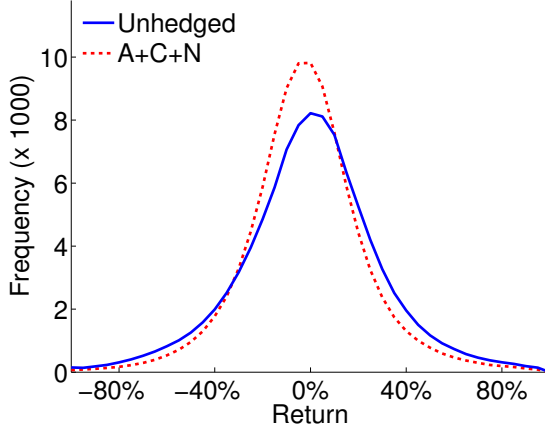


Figure 6.5: Returns 1997 to 2010

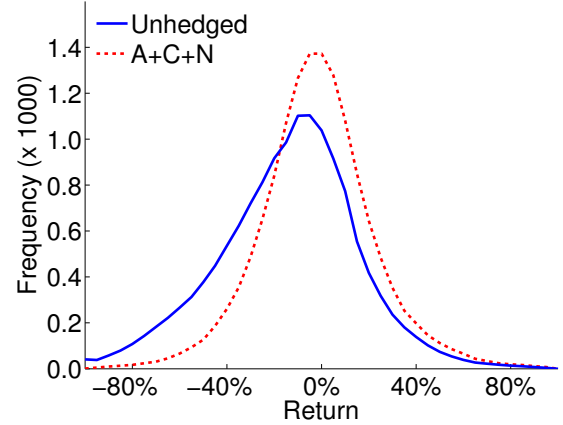


Figure 6.6: Returns during Crisis

using analyst cosine and news cosine values combined with correlation ( $A+C+N$ ), but the fact that hedging has more zero-centered returns is useful only during the crisis.

During this crisis period, we see that using a set of similar stocks does better than using a general market index (i.e., SPY). GICS generally does better in these measures than use of historical correlation (Correl) or news cosine values (News), but worse than analyst cosine values (Analyst). Using the Levene Test, the differences in the variances are significant in all cases, except GICS vs.  $A+C+N$  ( $p = 0.598$ ) and  $A+C+N$  vs. correlation ( $p = 0.068$ ). Analysts significantly outperform GICS ( $p = 0.005$ ) and  $A+C+N$  ( $p = 0.001$ ). From this perspective of variance, using Analysts does best at reducing risk. However, from the perspective of reducing the worst case losses,  $A+C+N$  does best as indicated by better results under the VaR[5], ES[5] and Roy risk measures. Analysts does better on DDown, but the other methods can be improved using the stop loss orders, as described earlier.

In summary, risk reduction improvements in a hedging scenario are observed over conventional hedging methods, such as using a market index or using an industry taxonomy to select the hedge stocks. These results further suggest strong value in the datasets and methods described in chapters 2 and 3.

### 6.3 Further Applications

In this section, we briefly describe applications areas where we believe further potential lies.

### 6.3.1 Comparative Analysis & Company Valuation

Valuation of companies is typically performed using one or a combination of the absolute and relative value models. We believe the research in this dissertation has implications for the improvement of relative value models, but to understand these models, we first describe absolute models.

With absolute models, the valuation is generally computed as the present value of expected future cash flows and does not incorporate market prices into model. Expected revenues and asset values are keys to these models. The advantages of these models are that when they are properly applied, they should be free of any biases in the market. The disadvantages are that they typically require a great deal of effort and expertise in financial statements since it is often necessary to make adjustments. For example, one-time asset sales, like the sale of a corporate office, should be removed from income since it cannot be expected to occur again. Further, the absolute approach inevitably requires many assumptions about specific values, such as future cost of capital, future sales growth, etc. Slight changes in these values can lead to dramatically different results, so absolute valuations can be considered “fragile” from this respect.

With relative models, the valuation is derived from observed market values for similar companies. A common approach in these valuations is to use a financial ratio computed over a company’s set of “peers.” For example, to compute a value for Kroger (a supermarket chain), the average price-to-earnings ratio of other supermarkets could be computed. Then, this ratio could be multiplied by the earnings of Kroger’s to find its value. Such an approach is often called a “comparables” approach and numerous other ratios, could be used, such as price-to-sales, price-earnings-growth (PEG), etc. Unlike absolute value models, these models are typically easier to compute, but also assume that the market is properly valuing the other companies. These approaches can also be useful in cases where not all financial information is available to perform a complete absolute valuation, such as with private companies. In the case of a private company, the comparables approach can be performed using data from public companies with the idea that this would reflect how the market would value the private company if it were public.

Key to the comparables approach is finding the right peer group. (Koller et al., 2005, pgs. 336-367) states

To analyze a company using comparables, you must first create a peer group. Most analysts start by examining the company's industry. But how do you define an industry? Sometimes, a company lists its competitors in its annual report. If the company doesn't disclose its competition, you can use an industry classification system such as Standard Industrial Classification (SIC) codes. [For example,] Home Depot's SIC code, however, contains more than 20 companies, many of which are not directly comparable because they sell very different products or rely on different business models. A slightly better but proprietary system is the Global Industry Classification Standard (GICS) system,...

This notion of selecting a peer group is precisely what the use of correlation, analysts and news helps to overcome. Just as with the long position hedging scenario in section 6.2, the pairwise values from chapter 3 can be used to select the most similar stocks to a given company. This approach may help with the problems associated with using industry groups, as identified in the snippet above. Whereas industry groups do not suggest which stocks within the group are most related to a given stock, the pairwise values provide an ordinal ranking of similarity.

Surprisingly little research has been performed in this area. Alford (1992) is one of the first to consider how the selection of the peer group might influence valuation. Bhojraj and Lee (2002) suggest a "warranted multiple" approach to selecting peers driven by accounting values. Henschke and Homburg (2009) provide an approach to repair valuations based on peer group differences. Given that essentially the only alternative to industry groups (or company guidance in annual reports) is to use accounting values, we believe there is much to be explored using other data sources, such as analysts, news and correlation.

### **6.3.2 Equity-Neutral Trading**

Section 2.1.1 mentioned cointegration could be used as an alternative objective to correlation, the focus of much of this dissertation. While correlation and cointegration are

mathematically different concepts, we believe the methods and datasets described in this dissertation may help to identify candidates for cointegration testing. Stocks that are cointegrated can be expected to revert to a mean spread when they divert. Thus, these stocks are excellent contenders for pairs trading and other forms of equity-neutral trading that exploit these concepts.

## Chapter 7

### Conclusion

The financial industry continually seeks higher efficiency, more automation and improved risk management. We believe this dissertation makes contributions in each of these areas. In chapter 3, the task of predicting future stock similarity (as measured through stock-return correlation) is considered. Two datasets, which previously had not truly been considered for this purpose, are demonstrated to have predictive power over future similarity, particularly at the highest levels of similarity. The first dataset is analyst coverage, where it is shown that two stocks that tend to be covered by the same analysts also tend to be similar. The second dataset is news articles, where it is shown that similar companies frequently co-occur in news articles. Finally, a method to combine analysts and news with historical correlation (A+C+N) is implemented and is shown to generally have greater predictive power than any of the input datasets alone.

In chapter 4, a hypergraph partitioning algorithm is presented that produces high quality partitions for imbalanced datasets. We believe this algorithm can be applied to a variety of domains, but we focus on its application to creating stock groups in chapter 5. The analyst and news datasets are each used again to form hypergraphs, where companies are represented as vertices and analysts or news articles are used to create edges. The goal is to partition the hypergraph such that, in the case of analysts, the number of instances where an analyst covers stocks in multiple parts of the partition is minimized, or, in the case of news, the number of articles containing companies in multiple parts is minimized.

This dissertation also offers an approach to forming groups that uses the pairwise values introduced in chapter 3. This approach has the advantage that it is easier to combine datasets, especially correlation. In forming stock groups, the A+C+N combination is again found to perform best, even having higher quality groups than a leading commercial industry



classification system, GICS.

In chapter 6, scenarios for two real world tasks are considered: diversification and hedging. In both cases, it is shown that the methods developed in previous chapters make improvements over traditional methods.

Prediction of company similarity is a vital facet of successful investment strategies and good financial analysis. This dissertation’s methods make steps towards the automation of these tasks as well as improving their predictive accuracy. We believe these methods should be considered by financial professionals to enhance or replace current processes that require strong prediction of company similarity.

## **7.1 Future Work**

### **7.1.1 Extraction of Relationships from News or Other Textual Data**

While the news dataset presented in this dissertation had weakest performance compared to analysts or correlation, news also has the greatest potential for improvements. This dissertation counted simple co-occurrences, which is a blunt method. Greater precision might be achieved by being more selective in the articles used. For example, it might be useful to avoid “business digests,” “yesterday’s gainers and losers” and other articles that are really a compendium of news events rather than a single story with related companies. Another useful heuristic might be to consider the proximity of the company names in the article, as suggested by Jin et al. (2012). Beyond heuristics such as these, natural language processing techniques could be developed to extract the actual relationship, such as competitor, supply chain partner, etc. Finally, more types of textual data, such as blog posts, company prospectuses, etc., could be used in addition to news articles.

### **7.1.2 Measure Timeliness**

Suppose a company enters a new market, such as Google’s recent entry into mobile phone development. Which of the similarity predictors suggested in this dissertation would react most quickly to this change? The efficient market hypothesis suggests that the stock market should react immediately. However, some amount of history is required to have a long

enough time series of returns such that correlation can be computed. So, correlation using historical returns might lag changes in company, depending on how it is calculated. Intuition suggests news should also react quickly since current events are of primary importance to news publications. On the other hand, analysts might be slow to adapt since it may take time for the research firm to decide to reassign a company to a different analyst. There are significant switching costs for the firm to change analysts since the new analyst will likely need to increase his/her knowledge of the company to be able to make earnings estimates and recommendations. Additionally, any brokerage clients familiar with the previous analyst will now need to develop a relationship with the new analyst. Finally, intuition also suggests that industry taxonomies might be slow to adapt to changes. GICS claims to be “evolving” (MSCI / Standard & Poor’s, 2002) by conducting reviews of company assignments at least annually and reviews of its taxonomical structure annually (Standard & Poor’s, 2008). Still, one might expect the time to conduct a review and publish changes to be longer than the time taken for several news articles to be published. Conducting experiments to measure the timeliness of each predictor would help to understand their “reaction times” and to incorporate this into prediction models.

Furthermore, it is well-established that correlations tend to differ depending on the time scale used in the correlation computation (Kinlaw et al., 2014). This dissertation focused on yearly correlations computed from daily stock returns. Results might differ using either less frequent returns, such as daily or monthly returns, or more frequent returns, such as intra-day returns. So, depending on the application, it may be important to determine if the relationships found in this thesis also exist at different time scales.

### **7.1.3 Labeling Stock Groups**

In chapter 5.2.1, we remarked that an advantage to industry taxonomies is that they apply a label to their groups. For instance, McDonald’s, Wendy’s and Panera might all be in the “restaurants” industry. This label helps the user to understand the business of the constituent companies. The automated grouping approaches suggested in this dissertation do not provide such labels by themselves. However, we believe it might be possible to

develop methods to automatically produce such labels by applying natural language processing techniques to company descriptions or other textual data regarding the companies in a group. For example, using the most frequent words in the descriptions of a group's companies might be useful for producing a label.

#### **7.1.4 Hypergraph Partitioning Improvements**

As suggested in section 4.3.4, multi-level hypergraph partitioners have been shown to have much shorter run times. A flat partitioner was used in this dissertation due to its simplicity, but using a multi-level framework in conjunction with the entropy-constraint and discount-cut concepts presented in this thesis might lead to improved run times and an ability to work with larger hypergraphs while still producing high quality partitions, particularly for imbalanced datasets.

## Bibliography

- Carol Alexander. *Market models: A guide to financial data analysis*. Wiley, Chichester, UK, 2001. ISBN 0471899755.
- Andrew W. Alford. The effect of the set of comparable firms on the accuracy of the price-earnings valuation method. *Journal of Accounting Research*, 30(1):94–108, 1992.
- Charles J Alpert and Andrew B Kahng. Recent directions in netlist partitioning: A survey. *Integration, the VLSI Journal*, 19:1–81, 1995.
- Donald R. Arbuckle. 1997 North American Industry Classification System – 1987 Standard Industrial Classification Replacement. *Federal Register*, 63(149):41696–41700, 1998.
- David Aronson. *Evidence-based technical analysis: Applying the scientific method and statistical inference to trading signals*. John Wiley & Sons, Hoboken, NJ, 2007.
- Cliff Asness, Andrea Frazzini, and Lasse H Pedersen. Leverage aversion and risk parity. *Financial Analysts Journal*, 68(1):47–59, 2012.
- K. Bache and M. Lichman. UCI machine learning repository, 2013. URL <http://archive.ics.uci.edu/ml>.
- Wolfgang Banzhaf, Peter Nordin, Robert E. Keller, and Frank D. Francone. *Genetic programming: An introduction on the automatic evolution of computer programs and its applications*. Morgan Kaufmann Publishers Dpunkt-verlag, San Francisco, Calif. Heidelberg, 1998.
- Shenghua Bao, Rui Li, Yong Yu, and Yunbo Cao. Competitor mining with the web. *IEEE Transactions on Knowledge and Data Engineering*, 20(10):1297–1310, October 2008.

- Brad M. Barber, Reuven Lehavy, and Brett Trueman. Comparing the stock recommendation performance of investment banks and independent research firms. *Journal of Financial Economics*, 85(2):490–517, 2007.
- Brad M. Barber, Reuven Lehavy, and Brett Trueman. Ratings changes, ratings levels, and the predictive value of analysts recommendations. *Financial Management*, 39(2):533–553, 2010.
- A. Bernstein, S. Clearwater, S. Hill, C. Perlich, and F. Provost. Discovering knowledge from relational data extracted from business news. In *SIGKDD Workshop on Multi-Relational Data Mining*, 2002.
- Sanjeev Bhojraj and Charles M. C. Lee. Who is my peer? a valuation-based approach to the selection of comparable firms. *Journal of Accounting Research*, 40(2):407–439, 2002.
- Sanjeev Bhojraj, Charles M. C. Lee, and Derek K. Oler. What’s my line? a comparison of industry classification schemes for capital market research. *Journal of Accounting Research*, 41(5):745–774, 2003.
- Mark Carlson. A brief history of the 1987 stock market crash with a discussion of the Federal Reserve response. *Board of Governors of the Federal Reserve System Finance and Economics Discussion Series #2007-13*, 2007.
- Stefano Cavaglia, Christopher Brightman, and Michael Aked. The increasing importance of industry factors. *Financial Analysts Journal*, 56(5):41–54, 2000.
- Ümit V. Çatalyürek and Cevdet Aykanat. PaToH: Partitioning tool for hypergraphs. <http://bmi.osu.edu/~umit/PaToH/manual.pdf>, 1999.
- Louis K.C. Chan, Josef Lakonishok, and Bhaskaran Swaminathan. Industry classification and return comovement. *Financial Analysts Journal*, 63(6):56–70, 2007.
- P.K. Chan, M.D.F. Schlag, and J.Y. Zien. Spectral k-way ratio-cut partitioning and clustering. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, 13(9):1088–1096, Sep 1994.

- Wesley S. Chan. Stock price reaction to news and no-news: Drift and reversal after headlines. *Journal of Financial Economics*, 70(2):223–260, 2003.
- Tarun Chordia, Richard Roll, and Avanidhar Subrahmanyam. Recent trends in trading activity and market quality. *Journal of Financial Economics*, 101(2):243–263, 2011.
- Pierre Cizeau, Marc Potters, and Jean-Philippe Bouchaud. Correlation structure of extreme stock returns. *Quantitative Finance*, 1(2):217–222, 2001.
- Richard N. Clarke. SICs as delineators of economic markets. *The Journal of Business*, 62(1):17–31, 1989.
- Roger Clarke, Harindra de Silva, and Steven Thorley. Risk parity, maximum diversification, and minimum variance: An analytic perspective. *The Journal of Portfolio Management*, 39(3):39–53, 2013.
- Jason Cong and Sung Kyu Lim. Multiway partitioning with pairwise movement. In *IEEE/ACM International Conference on Computer-Aided Design*, pages 512–516, Nov 1998.
- Alfred Cowles, III. Can stock market forecasters forecast? *Econometrica*, 1(3):309–324, 1933.
- James J. Cramer. *Jim Cramer’s real money: Sane investing in an insane world*. Simon & Schuster, New York, 2005.
- K.A.J. Doherty, R.G. Adams, N. Davey, and W. Pensuwon. Hierarchical topological clustering learns stock market sectors. In *ICSC Congress on Computational Intelligence Methods and Applications*, 2005.
- Robert J. Elliott, John Van Der Hoek, and William P. Malcolm. Pairs trading. *Quantitative Finance*, 5(3):271–276, 2005.
- Stuart Elliott. Why an agency said no to Wal-Mart. *New York Times*, page C1, Dec 15 2006.

- E.J. Elton, M.J. Gruber, and C.R. Blake. Survivor bias and mutual fund performance. *Review of Financial Studies*, 9(4):1097–1120, 1996.
- Eugene F. Fama and Kenneth R. French. Industry costs of equity. *Journal of Financial Economics*, 43(2):153–193, 1997.
- Lily Fang and Joel Peress. Media coverage and the cross-section of stock returns. *The Journal of Finance*, 64(5):2023–2052, 2009.
- C.M. Fiduccia and R.M. Mattheyses. A linear-time heuristic for improving network partitions. In *19th Conference on Design Automation*, pages 175–181, 1982.
- Kenneth French. Kenneth R. French-Data Library. [http://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data\\_library.html](http://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html), 2012.
- Michael Garey. *Computers and intractability: A guide to the theory of NP-completeness*. W.H. Freeman, San Francisco, 1979.
- Evan Gatev, William N. Goetzmann, and K. Geert Rouwenhorst. Pairs trading: Performance of a relative-value arbitrage rule. *Review of Financial Studies*, 19(3):797–827, 2006.
- Martin Gavrilov, Dragomir Anguelov, Piotr Indyk, and Rajeev Motwani. Mining the stock market: Which measure is best? In *Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '00, pages 487–496, New York, NY, USA, 2000. ACM.
- Charles Geisst. *Wall Street: A history*. Oxford University Press, Oxford New York, 2012.
- Joydeep Ghosh and Ayan Acharya. Cluster ensembles. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 1(4):305–315, 2011.
- David A Guenther and Andrew J Rosman. Differences between Compustat and CRSP SIC codes and related effects on research. *Journal of Accounting and Economics*, 18(1):115–128, 1994.

- Michael Hagenau, Michael Liebmann, Markus Hedwig, and Dirk Neumann. Automated news reading: Stock price prediction based on financial news using context-specific features. In *Proceedings of the 2012 45th Hawaii International Conference on System Sciences*, pages 1040–1049, 2012.
- Terrence Hendershott, Charles M. Jones, and Albert J. Menkveld. Does algorithmic trading improve liquidity? *The Journal of Finance*, 66(1):1–33, 2011.
- Stefan Henschke and Carsten Homburg. Equity valuation using multiples: Controlling for differences between firms, 2009. URL <http://ssrn.com/abstract=1270812>. Working Paper.
- Jeffrey Hooke. *Security Analysis and Business Valuation on Wall Street*. John Wiley & Sons, Hoboken, NJ, 2010.
- Changjian Hu, Liqin Xu, Guoyang Shen, and Toshikazu Fukushima. Temporal company relation mining from the web. In Qing Li, Ling Feng, Jian Pei, SeanX. Wang, Xiaofang Zhou, and Qiao-Ming Zhu, editors, *Advances in Data and Web Management*, volume 5446 of *Lecture Notes in Computer Science*, pages 392–403. Springer Berlin Heidelberg, 2009.
- Paul Jaccard. Étude comparative de la distribution florale dans une portion des alpes et du jura. *Bulletin de la Socit Vaudoise des Sciences Naturelles*, 37:547–579, 1901.
- Yingzi Jin, Ching-Yung Lin, Yutaka Matsuo, and Mitsuru Ishizuka. Mining dynamic social networks from public news articles for company value prediction. *Social Network Analysis and Mining*, 2(3):217–228, 2012.
- Charles Jones. A century of stock market liquidity and trading costs. Working paper, 2002.
- George Karypis and Vipin Kumar. Parallel multilevel k-way partitioning scheme for irregular graphs. In *Proceedings of the 1996 ACM/IEEE Conference on Supercomputing*, 1996.
- George Karypis and Vipin Kumar. hMETIS, a hypergraph partitioning package, version 1.5.3. User Manual, Nov 1998.



- George Karypis and Vipin Kumar. Multilevel k-way hypergraph partitioning. *VLSI Design*, 11:285–300, 2000.
- George Karypis, Rajat Aggarwal, Vipin Kumar, and Shashi Shekhar. Multilevel hypergraph partitioning: Application in VLSI domain. In *Proceedings of the 34th Annual Design Automation Conference*, DAC '97, pages 526–529, New York, NY, USA, 1997. ACM.
- Sally Katzen. Economic classification policy committee: Standard Industrial Classification replacement the North American Industry Classification System proposed industry classification structure. *Federal Register*, 60(143):38436–38452, 1995.
- M. G. Kendall. A new measure of rank correlation. *Biometrika*, 30(1/2):81–93, 1938.
- B W Kernighan and S Lin. An efficient heuristic procedure for partitioning graphs. *Bell System Technical Journal*, 49(2):291–307, 1970.
- William B Kinlaw, Mark Kritzman, and David Turkington. The divergence of the high and low frequency estimation: Causes and consequences. 2014. URL <http://ssrn.com/abstract=2433227>. Working Paper.
- Peter Klibanoff, Owen Lamont, and Thierry A. Wizman. Investor reaction to salient news in closed-end country funds. *The Journal of Finance*, 53(2):673–699, 1998.
- William R. Knight. A computer method for calculating Kendall’s tau with ungrouped data. *Journal of the American Statistical Association*, 61(314):436–439, 1966.
- Tim Koller, Marc Goedhart, and David Wessels. *Valuation: Measuring and Managing the Value of Companies*. John Wiley & Sons, Hoboken, NJ, 2005.
- Nan Li and Desheng Dash Wu. Using text mining and sentiment analysis for online forums hotspot detection and forecast. *Decision Support Systems*, 48(2):354–368, 2010.
- Alexander Ljungqvist, Christopher Malloy, and Felicia Marston. Rewriting history. *The Journal of Finance*, 64(4):1935–1960, 2009.
- Zhongming Ma, Olivia R.L. Sheng, and Gautam Pant. Discovering company revenue relations from news: A network approach. *Decision Support Systems*, 47(4):408–414, 2009.

- Sébastien Maillard, Thierry Roncalli, and Jérôme Teïletche. The properties of equally weighted risk contribution portfolios. *The Journal of Portfolio Management*, 36(4):60–70, 2010.
- Maureen Maitland and David M. Blitzler. The Global Industry Classification Standard (GICS): An overview for Standard & Poor’s U.S. Sector Indices, 2002.
- R.N. Mantegna. Hierarchical structure in financial markets. *The European Physical Journal B - Condensed Matter and Complex Systems*, 11(1):193–197, 1999.
- Harry Markowitz. Portfolio selection. *The Journal of Finance*, 7(1):77–91, 1952.
- Marina Meilă. Comparing clusterings an information based distance. *Journal of Multivariate Analysis*, 98(5):873–895, 2007.
- Dennis Meyers. Walk forward with the XAU bond fund system. *Technical Analysis of Stocks & Commodities*, 15(5):199–205, 1997.
- S. Micciche’, F. Lillo, and R. N. Mantegna. Correlation based hierarchical clustering in financial time series. In C. Beck, G. Benedek, A. Rapisarda, and C. Tsallis, editors, *Complexity, Metastability and Nonextensivity*, pages 327–335, 2005.
- MSCI / Standard & Poor’s. GICS - Global Industry Classification Standard, 2002.
- Michael P. Murray. A drunk and her dog: An illustration of cointegration and error correction. *The American Statistician*, 48(1):37–39, 1994.
- New York Times. Business digest. *New York Times*, page C2, Apr 12 2006.
- A Ochiai. Zoogeographic studies on the soleoid fishes found in Japan and its neighbouring regions. *Bulletin of the Japanese Society for Scientific Fisheries*, 22:526–530, 1957.
- David A. Papa and Igor L. Markov. Hypergraph partitioning and clustering. In Teofilo F. Gonzalez, editor, *Handbook of Approximation Algorithms and Metaheuristics*. Chapman & Hall/CRC, Boca Raton, 2007.
- Tobias Preis, Dror Y Kenett, H Eugene Stanley, Dirk Helbing, and Eshel Ben-Jacob. Quantifying the behavior of stock correlations under market stress. *Scientific reports*, 2, 2012.

- Sundaresh Ramnath. Investor and analyst reactions to earnings announcements of related firms: An empirical analysis. *Journal of Accounting Research*, 40(5):1351–1376, 2002.
- Pierre-Alain Reigner, Romain Allez, and Jean-Philippe Bouchaud. Principal regression analysis and the index leverage effect. *Physica A: Statistical Mechanics and its Applications*, 390(17):3026–3035, 2011.
- Samuel Rönqvist and Peter Sarlin. From text to bank interrelation maps. *ArXiv e-prints*, (1306.3856), 2013. Working Paper.
- Arthur D. Roy. Safety first and the holding of assets. *Econometrica*, 20(3):431–449, 1952.
- L.A. Sanchis. Multiple-way network partitioning. *IEEE Transactions on Computers*, 38(1):62–81, Jan 1989.
- Jeremy Siegel. *Stocks for the long run*. McGraw-Hill, New York, 2008.
- Andrew Ross Sorkin and Steve Lohr. Procter closes \$57 billion deal to buy Gillette. *New York Times*, page A1, Jan 28 2005.
- Standard & Poor’s. Global Industry Classification Standard Methodology, 2008.
- Alexander Strehl and Joydeep Ghosh. Cluster ensembles — a knowledge reuse framework for combining multiple partitions. *The Journal of Machine Learning Research*, 3:583–617, 2003.
- Pang-Ning Tan, Michael Steinbach, and Vipin Kumar. *Introduction to Data Mining*. Pearson Addison Wesley, Boston, 2005.
- Taffee T. Tanimoto. IBM internal report. *Nov*, 17:1957, 1957.
- 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.
- Thomson Reuters. Estimates. [http://thomsonreuters.com/products/financial-risk/content/07\\_004/thomson-reuters-estimates-fact-sheet.pdf](http://thomsonreuters.com/products/financial-risk/content/07_004/thomson-reuters-estimates-fact-sheet.pdf), 2014. Brochure.

- A. Topchy, A.K. Jain, and W. Punch. Clustering ensembles: Models of consensus and weak partitions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(12):1866–1881, Dec 2005.
- Bence Tóth and János Kertész. Increasing market efficiency: Evolution of cross-correlations of stock returns. *Physica A: Statistical Mechanics and its Applications*, 360(2):505–515, 2006.
- U.S. Commodity Futures Trading Commission and U.S. Securities and Exchange Commission. Findings regarding the market events of May 6, 2010. *Report of the Staffs of the CFTC and SEC to the Joint Advisory Committee on Emerging Regulatory Issues*, 10, 2010.
- U.S. Securities and Exchange Commission. Beginners’ guide to asset allocation, diversification, and rebalancing. <http://www.sec.gov/investor/pubs/assetallocation.htm>, 2009.
- U.S. Securities and Exchange Commission. Analyzing analyst recommendations. <http://www.sec.gov/investor/pubs/analysts.htm>, 2010.
- James Valentine. *Best Practices for Equity Research Analysts: Essentials for Buy-Side and Sell-Side Analysts*. McGraw-Hill, New York, 2011.
- Maximilian A. M. Vermorken. GICS or ICB, how different is similar? *Journal of Asset Management*, 12(1):30–44, 2011.
- Yen-Chuen Wei and Chung-Kuan Cheng. Towards efficient hierarchical designs by ratio cut partitioning. In *IEEE International Conference on Computer-Aided Design*, pages 298–301, Nov 1989.
- Ivo Welch. Herding among security analysts. *Journal of Financial Economics*, 58(3):369–396, 2000.
- Wyatt Wells. Certificates and computers: The remaking of wall street, 1967 to 1971. *The Business History Review*, 74(2):193–235, 2000.

Rui Xu and Donald C. Wunsch. *Clustering*. Wiley, Oxford, 2009.

Li Yang, Francis Tapon, and Yiguo Sun. International correlations across stock markets and industries: Trends and patterns 1988-2002. *Applied Financial Economics*, 16(16): 1171–1183, 2006.

John Robert Yaros and Tomasz Imieliński. Crowdsourced stock clustering through equity analyst hypergraph partitioning. In *IEEE Conference on Computational Intelligence for Financial Engineering & Economics*, 2013a.

John Robert Yaros and Tomasz Imieliński. Imbalanced hypergraph partitioning and improvements for consensus clustering. In *IEEE Conference on Computational Intelligence for Financial Engineering & Economics*, 2013b.

John Robert Yaros and Tomasz Imieliński. A Monte Carlo measure to improve fairness in equity analyst evaluation. In *Applied Mathematics, Modeling and Computational Science*, 2013c.

John Robert Yaros and Tomasz Imieliński. Using equity analyst coverage to determine stock similarity. In *IEEE Conference on Computational Intelligence for Financial Engineering & Economics*, 2014a.

John Robert Yaros and Tomasz Imieliński. Data-driven methods for equity similarity prediction. *Quantitative Finance*, Special Issue on Big Data Analytics: Algorithmic Trading and Financial Text Mining, 2014b. (to appear).

John Robert Yaros and Tomasz Imieliński. Diversification improvements through news article co-occurrences. In *IEEE Conference on Computational Intelligence for Financial Engineering & Economics*, 2014c.

Wenbin Zhang and Steven Skiena. Trading strategies to exploit blog and news sentiment. In *International AAAI Conference on Weblogs and Social Media*, 2010.