Using Equity Analyst Coverage to Determine Stock Similarity

John Robert Yaros and Tomasz Imieliński
Department of Computer Science
Rutgers University
Piscataway, NJ 08854-8019
Email: {yaros,imielinski}@cs.rutgers.edu

Abstract—With the observation that equity analysts tend to cover similar stocks, we propose a simple, intuitive method to convert their coverage sets into pairwise similarity values among stocks. These values are shown to have a strong positive relationship with future stock-return correlation. Further, these values are easily combined with historical correlation. Together, they produce more accurate predictions of future correlation than either does separately. Using an agglomerative clusterer and a genetic algorithm in a pipeline approach, we use the pairwise values to form clusters of similar stocks. We compare these clusters against a leading industry classification system, GICS, finding that the clusters from the combined analyst and correlation pairwise values tend to perform at least as well as GICS and often better. In an application of our pairwise values, we consider a hypothetical scenario where an investor wishes to hedge a long position in a single stock. Our results indicate that using the analyst similarity values to select a hedge portfolio leads to greater risk reduction than using GICS or hedging with a broad-market index. Using a combination of historical correlation with the analyst values leads to even greater improvements.

I. INTRODUCTION

Sell-side equity research firms typically employ many analysts. To satisfy client demand for research on a large number of companies, the research firm must decide how to assign those companies to its analysts. If an analyst is assigned companies that are all in the same industry, then s/he can focus on understanding the dynamics and environment of that single industry. Otherwise, the analyst would need to divide efforts among industries. Moreover, for each company the analyst covers, that analyst must have strong knowledge of the company’s competitors to be able to effectively estimate earnings and make buy/hold/sell recommendations. So, if an analyst is covering a particular company, s/he might as well be covering the competitor companies in the same industry anyway. Thus, strong incentives exist for the research firm to assign its analysts to cover highly similar stocks.

This work presents a method that harnesses these coverage assignments to make strong predictions about future stock similarity, particularly as measured by correlation. As a preview, suppose there are 10 analysts covering both stock A and stock B, while 20 analysts cover at least one of stock A or B. Our computed similarity value is 10/20 = 0.5. The computation is simple, but remarkably powerful in quantifying similarity.

To make comparisons, we use two data sources that have traditionally been used to determine stock similarity: historical correlation and industry classification systems. We use the Global Industry Classification System (GICS) as the industry taxonomy since previous research [1] has indicated GICS is superior to other systems at grouping highly correlated stocks.

Our method produces a similarity value for each pair of stocks in a given universe. This pairwise structure is similar to correlation, which makes it easy to combine these two datasets to produce more robust similarity values. This pairwise structure can also be more effective in many applications than a rigid classification system like GICS. For example, 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), a real estate investment trust (REIT) that owns and operates hotels. 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. Thus, a good hedging portfolio for MAR would likely include HST. However, if a trader decided to hedge using the constituents of MAR’s GICS sector, it would not include HST. MAR is classified under Consumer Discretionary, while HST is under Financials because it is a REIT. The industry classification completely misses the relationship between MAR and HST because each company is forced into a single sector.

One may argue that more sophisticated models, such as the Barra US Equity Model version 4 (USE4), can account for such situations because each company is assigned weights across multiple sectors. So, MAR and HST would likely share a large weight in one, or possibly more, sectors. However, these weights still do not make clear the strength of individual relationships between companies. 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. These relationships are easily obscured by industry classifications, even those using multiple weights to different sectors or factors.

The aforementioned pairwise representation in our analyst similarity values conveys these relationships more easily. Whereas MAR and HST appear in completely separate GICS sectors, their analyst similarity value is 19/31, meaning that in the previous year, out of the 31 analysts that covered either MAR or HST, 19 covered both. Such a high value indicates high similarity between the companies. TJX and ROST have similarity value 15/22, while TJX and ANF have value 13/48.
On the other hand, a user who wishes to get a fast overview of the entire market may prefer a simple partition of stocks, such as that found in classification systems like GICS. For example, if a significant broad-based news event occurred, a trader may want to view the impact to her portfolio in sectors such that she can quickly identify which areas need to be examined most quickly for hedging or other actions. Accordingly, we offer a method using two well-known techniques, agglomerative clustering and genetic algorithms, to form stock groups from our pairwise values. This enables comparisons with GICS, providing a strong reference point for our method.

This article is composed as follows. We first discuss related work in section II. In section III, we present a method to compute the stock similarity values through a Jaccard index. We show these analyst Jaccard values are positively related to future correlation. We also show they can be easily be combined with historical correlation for an even stronger prediction of future correlation. In section IV we provide a method to partition the stock universe (i.e., form stock groups) from pairwise similarity values. Against GICS, we compare the partitions formed by the analyst Jaccard values, historical correlation and their combination. In section V we discuss potential applications and examine a particular hedging scenario in detail. In section VI we conclude.

II. RELATED WORK

Much research effort has gone into determining whether analysts are accurate at estimating earnings and recommendations (see [2], for example). Even commercial products, such as StarMine, are devoted to tracking and evaluating analyst performance. Yet, surprisingly little research has considered the use of analyst coverage to determine stock similarity. In past literature, we find [3] 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 [4] and [5] had found issues with the industry classification scheme that was predominate at the time, the Standard Industry Classification (SIC). The author 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 the author 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.

In [6], we used a method called hypergraph partitioning to form stock groups from analyst coverage. Using a performance measure described in section IV-B, these groups found to outperform an academic scheme from [7] and perform roughly on par with GICS. The study demonstrates that analyst coverage is highly useful in determining stock similarity. However, the financial practitioner is unlikely to find hypergraph partitioning readily understandable without knowledge of graph theory. Additionally, the hypergraph representation for coverage data is not easily combined with other data, such as historical correlations. Finally, the output of the hypergraph partitioning is a set of stock groups and, as mentioned in section I, stock groups are less well suited to applications where individual similarities between stocks are important.

TABLE I. HYPOTHETICAL ANALYST COVERAGE

<table>
<thead>
<tr>
<th>Analyst</th>
<th>Companies Covered</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alice</td>
<td>CMG, MCD, YUM</td>
</tr>
<tr>
<td>Bob</td>
<td>CMG, DRI, EAT, SBUX</td>
</tr>
<tr>
<td>Eve</td>
<td>DRI, EAT</td>
</tr>
</tbody>
</table>

TABLE II. SIMILARITY VALUES

<table>
<thead>
<tr>
<th></th>
<th>CMG</th>
<th>DRI</th>
<th>EAT</th>
<th>MCD</th>
<th>SBUX</th>
<th>YUM</th>
</tr>
</thead>
<tbody>
<tr>
<td>CMG</td>
<td>1/2</td>
<td>-</td>
<td>1/2</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>DRI</td>
<td>1/3</td>
<td>-</td>
<td>1/3</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>EAT</td>
<td>1/3</td>
<td>1</td>
<td>1/2</td>
<td>0</td>
<td>1/2</td>
<td>1/3</td>
</tr>
<tr>
<td>MCD</td>
<td>1/2</td>
<td>-</td>
<td>1/2</td>
<td>-</td>
<td>0</td>
<td>1/2</td>
</tr>
<tr>
<td>SBUX</td>
<td>1/2</td>
<td>1/2</td>
<td>1/2</td>
<td>1/2</td>
<td>0</td>
<td>1/3</td>
</tr>
<tr>
<td>YUM</td>
<td>1/2</td>
<td>1/3</td>
<td>1/3</td>
<td>1/3</td>
<td>0</td>
<td>-</td>
</tr>
</tbody>
</table>

III. SIMILARITY MEASURE

We measure the similarity between two stocks as the ratio of the count of analysts covering both stocks divided by the count of analysts covering either stock:

$$J_{ij} = \frac{|C_i \cap C_j|}{|C_i \cup C_j|}$$  \hspace{1cm} (1)

where $C_i$ represents the set of analysts covering a particular stock $i$ and $C_j$ represents the set of analysts covering a different stock $j$. For two sets of any objects, this ratio of intersection to union is commonly called the Jaccard index since it was originally published by Paul Jaccard [8].

Consider the hypothetical analyst coverage shown in Table I. The corresponding Jaccard values are shown in Table II. For example, the similarity between Chipotle Mexican Grill (CMG) and McDonald’s (MCD) is 1/2. One analyst, Alice, covers both CMG and MCD, so the numerator is 1. Two analysts, Alice and Bob, cover either CMG or MCD, so the denominator is 2. The minimum possible value for our analyst Jaccard is 0.0 indicating there are no analysts in common between the two stocks and, therefore, have no similarity. The maximum possible value is 1.0 indicating every analyst covering the first stock also covers the second stock, and vice versa. The “distance” between two stocks can be computed as $1 - J_{ij}$, and thus, stocks that are most dissimilar will be furthest from each other.

A. Positive Relationship with Future Correlation

Before we use these Jaccard values in applications, we consider whether they are predictive of future correlation. For this experiment and all others in this article, we use a dataset from I/B/E/S, which records earnings estimates and recommendations from major brokerages. We slice our data into individual years and consider an analyst to be covering a stock if the analyst made at least one earnings estimate for that stock in the past year. For stock returns, we use a dataset from the Center for Research in Security Prices (CRSP) which offers a daily return computation that includes splits, cash and stock dividends, and other distributions. For our universe of stocks, we use the broad market S&P 500 index, which is composed of the S&P 500 large-cap, S&P 400 mid-cap and S&P 600 small-cap stocks. We use Compustat, a product of S&P, to determine the composition of the S&P 500 each year from 1996 to 2010. In order to avoid survivorship bias [9], we fix the index compositions 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.
We wish to use the similarity values in a predictive sense. Accordingly, we use walk-forward testing \[^{[1]}\] \[^{[11]}\] \[^{[11]}\] \[^{[1]}\], whereby we measure for a relationship between the Jaccard values computed over one year’s data and the stock return correlations from the subsequent year (i.e., the future correlation). In Fig. 1 we group pairs of stocks into five separate ranges by their Jaccard index. As seen in the figure, a higher range of Jaccard values nearly always means higher future correlation. This effect is stable despite the general trend towards increased correlations - a trend also observed by others (e.g., \[^{[11]}\] \[^{[1]}\]).

In Fig. 2 we see that stocks with high Jaccard values are infrequent compared to low Jaccard values (note the logarithmic scale on the Y-Axis). In other words, most pairs of stocks have few, if any, analysts covering both. In fact, even though 99.6% of stocks in the S&P 1500 were covered by 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. As will be described in section III-B this means the analyst Jaccard is generally not effective in differentiating moderately and slightly similar stocks, but is effective at determining highly similar stocks (which is often most important).

To test the statistical significance of the relationship between Jaccard values and correlation, we use the nonparametric Kendall’s tau, 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 Jaccard has higher future correlation. Discordance means the pair with higher Jaccard has lower future correlation. Kendall’s tau counts and normalizes all possible occurrences to output a number between −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 III shows the Kendall’s tau values and corresponding odds-ratios when testing for a relationship between the analyst Jaccard and future correlation. In each year, the Kendall’s tau values are greater than zero with statistical significance well below 0.1%. So, analyst Jaccard values are clearly positively related to future correlation.

B. Ease of Combination with Historical Correlation

The similar structures of historical pairwise correlation and our analyst Jaccard values make it possible to easily combine them into a single set of pairwise values. For example, just as the matrix of Table II was used to display the analyst Jaccard values, an analogous matrix can be composed of pairwise correlation values.

Suppose we wish to better predict correlation in the next year. Let \(C\) denote the matrix of the past year’s correlation values and let \(A\) denote the matrix of pairwise analyst Jaccard values. We can combine them by

\[
H = (1 - \alpha)C + \alpha A
\]

where \(\alpha\) is a coefficient of strength between \(C\) and \(A\). We calculate \(\alpha\) by estimating the value that would have been most predictive in the previous year\[^{[1]}\]. That is, if we wish to calculate a value of \(\alpha\) to use in predicting correlation for year 2005, we regress the correlation values \(C\) of 2004 against the 2003 correlation \(C\) and Jaccard values \(A\). This finds the \(\alpha\) that would have best predicted the future (2004) correlation from the previous (2003) correlation and Jaccard values. Then, to predict for 2005, we will use that \(\alpha\) value to combine \(C\) and \(A\) with 2004 data.

Figs. 3 & 4 depict the predictive strength of the analyst Jaccard values, historical correlation and the Jaccard & correlation combination (corresponding to \(A, C\) and \(H\), respectively, in equation 2). We quantify the predictive strength with a “hit rate,” which we will explain through an example. Suppose we wish to predict the top five most similar stocks for Walmart. We rank all other stocks using one of the measures, such as the analyst Jaccard, on the previous year’s data. We then choose the top five most similar: 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 \(3/5 = 0.6\). Figs. 3 & 4 display the average hit rate of all stocks in the S&P 1500 where at least one analyst covered the stock in the previous year. Fig. 3 displays the top 5 hit rate, while Fig. 4 displays the top 50 hit rate.

As the figure indicates, the relative performance of the analyst Jaccard degrades from selecting the top 5 to selecting

\[^{1}\text{Correlation values range from } -1 \text{ to } 1, \text{ whereas analyst Jaccard values range from } 0 \text{ to } 1. \text{ Hence, the values are slightly different. However, one can normalize the values by subtracting the mean and dividing by standard deviation. We perform this adjustment in our combinations, although it may not be necessary in practice since } \alpha \text{ will adjust to differences in magnitudes and correlations between stocks are rarely negative anyway.}

\[^{2}\text{We use least-squares regression with constraint that } 0 \leq \alpha \leq 1.\]
the top 50. As mentioned in section III-A, the similarity value for most pairs of stocks is zero because there are no analysts between them. For example, the similarity between two clothing retailers is well determined by the analysts, but the similarity between a clothing retailer and an automotive producer will likely be zero, just as the similarity between a clothing retailer and a gas utility will likely be zero. This means that the analyst Jaccard is not great at differentiating which stocks are slightly or moderately related, but is much better at determining which stocks are most related to a given stock. As we will see in section IV, the most similar stocks are often the most important, particularly in hedging.

Also evident in Figs. 3 & 4 is that the combination of analyst Jaccard and historical correlation generally has best performance. Even with the top 50 where analyst Jaccard performance is relatively lower, combining the analyst Jaccard values with historical correlation boosts accuracy over just using each individually. In section IV we will see that this combination can be form highly correlated stock groups. In section IV we will see it can be used to improve hedging.

IV. STOCK GROUPS

In [6], we used hypergraph partitioning over analyst coverage to form sector groups with quality roughly on par with GICS. We also used an “off-the-shelf” algorithm to form groups from the previous year’s correlations. The historical groups generally underperformed both the analyst groups and GICS. In this section, we revisit this task. We first describe a method over pairwise distances that is better suited for determining stock groups than the previously used “off-the-shelf” algorithm. Second, we apply this method to historical correlations and to the analyst data and draw comparisons with GICS. Finally, we demonstrate how the historical correlations can be combined with analyst data to do even better.

We use the analyst data here to form groups in order to be able to draw comparisons with GICS, but we emphasize that the analyst Jaccard values themselves (perhaps also in combination with historical correlation) can be more powerful in many applications due to their pairwise nature. We provide an example of such an application in section IV.

A. Global Industry Classification System (GICS)

We briefly describe GICS since it serves as a main point of comparison. GICS is maintained by S&P and MSCI. As depicted in Fig. 5, it is composed of four levels of granularity. At the least granular level there are sectors. These are subdivided into industry groups, then industries and, finally, sub-industries. The count of groups at each level of granularity may evolve over time to match changes in the economy, but

at the start of 2010, the final year in our dataset, there were 10 sectors, 24 industry groups, 68 industries and 154 sub-industries. As explained in their brochure, MSCI and S&P “jointly assign a company to a single GICS sub-industry according to the definition of its principal business activity as determined by the two companies. Revenues are a significant factor in determining principal business activity; however, earnings and market perception are also important criteria for classification.” [13]

To contrast GICS and the use of analyst coverage, note that GICS assignments are controlled by only two companies, S&P and MSCI, whereas our analyst coverage method essentially “crowdsources” opinions across many stock research firms. Moreover, GICS, by design, forces each stock into a single group (i.e., sector, industry, etc.). This means that if there are highly similar stocks that happen to fall into separate groups (as in the MAR and HST example in section IV), GICS will have no indication of that similarity. Finally, GICS does not provide any information about the strength of individual relationships within a group (as in the apparel retail example in section IV).

B. Quality Measures

The quality measure used to compare stock groupings in [6] originated in a study comparing several industry classification systems [1]. 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

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

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 in [6] to allow stocks that 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:

$$
\psi = \frac{\sum_{I \in 1} (\bar{\rho}_I - \bar{\phi}_I)}{||I||} \quad \theta = \frac{\sum_{I \in 1} |I| \cdot (\bar{\rho}_I - \bar{\phi}_I)}{\sum_{I \in 1} |I|}
$$

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

Our previous study [6] used a simple agglomerative clusterer based on a distance measure of one minus correlation. In the clustering algorithm, each stock begins as its own group. The algorithm then proceeds to combine groups by selecting two groups which have lowest distance between the furthest pair of stocks between the two groups. This is known as complete linkage. The algorithm is effective in many scenarios, but can easily produce a less than optimal set of groups. Instead, we propose to use an agglomerative clusterer that directly attempts to optimize for the grouping quality measure. That is, we attempt to find the optimal stock grouping such that $\theta$ is maximized.

However, optimizing with 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 group. The large group has lower intra-group correlations ($\bar{\rho}_I$), but the inter-group correlations ($\bar{\phi}_I$) are much lower, leading to larger $\psi$ and $\theta$ measures. Consider Table IV. 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.

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 $\phi_I$) [14]. 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 a suboptimal group. Therefore, we suggest $\psi$ and $\theta$ be replaced by

$$
\kappa = \frac{\sum_{I \in 1} \rho_I}{||I||} \quad \gamma = \frac{\sum_{I \in 1} |I| \cdot \rho_I}{\sum_{I \in 1} |I|}
$$

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 article.

C. Forming Groups from Correlation and Analyst Data

Finding the optimal set of stock groups maximizing $\gamma$ (or $\kappa$) is computationally difficult. 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 $S(n)$ denotes Stirling numbers of the second kind. An exhaustive search is not feasible. However, 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 into a single group. 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 all $2.76 \times 10^{1493}$ combinations 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 4 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 were $N = 1000$, $M = 2$, $P = 25\%$, $Q = 0.5\%$ and $T = 50$. Genetic algorithms have been shown to be effective in a variety of settings (see [15] for an introduction), and we have found ours does well in improving the initial solution produced by the agglomerative clusterer.

<table>
<thead>
<tr>
<th>TABLE IV. HYPOTHETICAL CORRELATIONS</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
</tr>
<tr>
<td>---</td>
</tr>
<tr>
<td>-</td>
</tr>
<tr>
<td>2</td>
</tr>
<tr>
<td>3</td>
</tr>
<tr>
<td>4</td>
</tr>
<tr>
<td>5</td>
</tr>
<tr>
<td>6</td>
</tr>
<tr>
<td>7</td>
</tr>
<tr>
<td>8</td>
</tr>
<tr>
<td>9</td>
</tr>
</tbody>
</table>
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. As mentioned in section III-B, the pairwise analyst Jaccard values have similar structure to the pairwise correlation data. In fact, we use the same algorithms on the analyst Jaccard values with the same quality measure, \( \gamma \), except replacing pairwise correlation values \( \rho_{ij} \) with the pairwise analyst Jaccard values \( J_{ij} \). In a similar fashion, we use the values from the combined analyst Jaccard and historical correlation method of section III-B to generate another comparative set of groups. The workflow for the entire process with the combination method is shown in Fig. 6.

\[ \text{Start} \]
Want prediction for year T
Compute pairwise correlations, \( C_2 \) for year T-2, and \( C_1 \) for year T-1

Find \( \alpha \) that is most predictive from year T-2 to T-1, by constrained least squares linear regression
\[ C_2 = C_1 + \alpha A_1 \]

Using data from year T-1, Calculate \( H_1 = C_1 + \alpha A_1 \)

Run Agglomerative Cluster on \( H_1 \) to form candidate stock groups, \( \hat{G} \)

Using \( \hat{G} \) as initialization, run Genetic Algorithm on \( H_1 \) to form candidate stock groups, \( G \)

End
Return \( G \) as prediction for year T

Fig. 6. Workflow to create stock groups from combined historical correlation and analyst Jaccard values

D. Results

A comparison of each of the methods at each GICS level is shown in Figs. 7, 8, 9 & 10. Because correlations change dramatically from year to year, we do not plot the intra-group correlation values (\( \gamma \) in section IV-B) 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 IV-C 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.

As seen in the figures, historical correlation generally performs slightly worse than GICS. These results are similar to those found in previous studies [1]. However, since the algorithms of section IV-C are directly optimizing the quality measure rather than using off-the-shelf algorithms, the results are better than the previous studies would indicate. In fact, correlation’s underperformance versus GICS is only statistically significatn at the 5% level under a paired t-test for GICS’s sub-industry granularity (\( p = 0.002 \)). All other granularities are not statistically significant.

The analyst Jaccard performs most poorly when the number of groups is lowest. At the Sector and Industry Group levels, it generally underperforms both GICS and correlation, although still much better than random (as a zero on the y-axis would indicate). When more groups are involved, it performs much better. At sub-industry granularity, analysts significantly outperform both GICS (\( p = 6.3 \times 10^{-4} \)) and correlation (\( p = 3.5 \times 10^{-2} \)).

The analyst Jaccard and correlation combination method does best. At sector granularity, none of GICS, correlation or the combination method are statistically better than any of the others. At industry-group granularity, the combination method is not statistically better than GICS (\( p = 0.091 \)), but is better than correlation (\( p = 0.006 \)). At both industry and sub-industry granularities, the combination method statistically outperforms GICS, correlation and the analyst Jaccards (\( p < 10^{-4} \) for each). Therefore, we see that the combination method generally performs at least as well as its component parts, correlation and the analyst Jaccard. We also see that it generally performs at least as well as GICS. In most cases, it does much better.

V. Application to Hedging

Recall that the pairwise representation of stock similarity provides many advantages over groups. As discussed in section IV, TJX Companies (TJX), Ross Stores (ROST) and Abercrombie & Fitch (ANF) all fall into the apparel retail group, however, TJX and ROST, both discount retailers, are intuitively more similar to each other than to ANF, a higher-end
retailer. Hence, one potential application of pairwise similarity is simply to allow an investor or researcher to find the most similar stocks to a given stock. This can help, for example, an investor with a view on a particular mining company to analyze its competitors and possibly also find an alternative stock in the same space. They can also be used for risk management since one can expect similar stocks to have commonalities in their time series of returns. Conversely, one can improve portfolio diversification by avoiding similar stocks (e.g., ensure no two stocks in the portfolio have a non-zero analyst Jaccard value).

For the remainder of this article, we will explore the application of our pairwise similarity values (analyst Jaccard, historical correlation and their combination) to hedging. We consider 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 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 has 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 analyst Jaccard values, correlation and their combination. We compare these hedging measures 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.

A. 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 2010 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. 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 Jaccard values, or the optimal combination thereof (as described in section III-B). 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 500 trading days (roughly two calendar years) prior to the run’s start date:

$$r_{s,t} = \alpha + \beta \cdot r_{h,t} + \epsilon_t$$

(3)

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.

B. Results

Table [V] contains results over the 100,000 runs for the entire 1997-2010 time period. Table [VI] 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 absolute 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: the probability that returns are less than X%. (-25%, -50%, -75% & -100% shown.)

As seen in Table[V], 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.

Still, when viewing the average return (AvgRet) in Table[V], the reduction in risk may not seem worth the reduction in return. Stock values tend to rise in general [16] 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 VI. Both the risk factors are reduced and the average return (AvgRet) is higher with hedging. The difference is further evident in the histograms of Figs. [1] and [2], where the variance (i.e., risk) is lower in both periods for hedging using analyst Jaccard values combined with correlation (Ana+Cor), but the fact that hedging has more zero-centered returns is useful only during the crisis.

Among the hedge methods, we see that using a set of similar stocks does better than using a general market index (i.e., SPY). GICS does better in these measures than use of historical correlation (Correl), but worse than analyst Jaccard values (Analyst) or the combination method (Ana+Cor), which does best. These results are consistent with the findings in section III-B where the combination method did best at predicting future correlation. Using the Levene Test, the difference in variances during the crisis in each pair of methods is significant at the 5% level, except in two pairs: Correl outperforming GICS (p = 0.399), and Ana+Cor outperforming Analysts (p = 0.094).

VI. SUMMARY

Due in part to competition’s constant drive towards efficiency, sell-side equity research analysts tend to cover highly related stocks. Harnessing this fact, we describe a simple method to quantify the similarity between two stocks by computing a Jaccard index of analyst coverage. We demonstrate that the value of the Jaccard index is positively related to the future pairwise correlation between the two stocks. In addition, we show that historical correlation and analyst Jaccard values can easily be combined to find values that are more predictive of future correlation than either is separately.

We describe a pipeline approach to constructing stock groups from pairwise similarity values. Using the analyst Jaccard values, historical correlation, and the combination of both, we compare our groups against a leading industry classification system, GICS. We find that the combined analyst and correlation values produce groups with quality that is generally at least as good as GICS and often better, especially as the number of groups increases, such as with sub-industries.

**REFERENCES**


