2016 Goldman Sachs Quant Quest

DoWon Kim, Yash Patel, Jennifer Yin, Patrick Zeng

Princeton University

1 Abstract

This study examines, quantifies, and interprets in an economic context the magnitude of intercompany correlation in between S&P500 companies using collaboratively contributed content on each of the companys Wikipedia webpages, along with accessory Wikipedia pages of relevance to these companies. In our approach, we use a combination of a proprietary Manual Covariance Analysis (MCA) and Latent Dirichlet Allocation (LDA) to generate a 500 by 500 matrix of one-directional relational magnitudes from each company to each other company.

2 General Approach

Relationships between companies are often times multi-dimensional and complex in nature; consequently, the task of simplifying the relationship between two companies down to a single quantitative magnitude requires careful definition of what it means to be correlated. Our definition of correlations between companies is motivated by their end economic application; while it may be more straightforward to define correlation quantitatively based on a limited number of measurable factors, such an academically rigorous definition would have limited practicality economically. Instead, we use a multi-pronged approach to capture both factor-based and content-based indicators of correlation, as this better captures the multifaceted economic relationships between companies in the real world.

What do we foresee as the ultimate application of our correlatory matrix? The correlations we establish between companies should give us a perspective on which related companies to assess when determining one company's financial movement. In the reverse direction, we also seek to have a measure of how significantly each of the companies on the S&P500 would be affected given a substantial change in one company's financial performance. In addition, our correlations between companies may differ depending on direction. For example, a manufacturer of one component for a tech product may be heavily correlated with the tech product's company, but the tech company may be minimally correlated with the manufacturer, as the manufacturer may be replaceable.

3 Technical Approach

3.1 Factor-Based Analysis

Motivated by this end goal, we choose to utilize a combined factor-based and content-based approach as described above. In our factor-based analysis, we examine key factors of companies we know to be economically significant, including company size, market sector, geographical location served, net income, and total assets. Weighting these components individually based on their contributing significance to similarities between companies, we then use MCA to determine a measurement of correlation based on these factors. Important to note is that this measure of correlation between companies is symmetrical, as the factors compared between each pair of companies hold the same value in either direction. This factor-based

analysis forms an economic backbone for the correlations between companies.

3.2 Content-Based Analysis

Our content-based analysis uses LDA to analyze the remaining content of the company's Wikipedia page. In this part, the matrix becomes asymmetric because each individual page now includes many references to other companies that do not align the same way that something like industry does. There is far more variation and therefore changes the values for different companies asymmetrically as they should be.

4 Results and Discussion

4.1 Economic Interpretation

Every cell in our vector is a number less than one representing the normalized magnitude of correlation between two companies in the S&P500. This serves first and foremost as a guide for which company to qualitatively assess when trying to decide how a company will move in the future. Given the proper technology, this would also serve as a guide for which companies to follow most closely, and how to react sudden changes in their stock price. For example, we found that QCOM had a high correlation factor for AAPL. In this specific case, we can see that this is because QCOM produces one of AAPL's parts which should give it a positive correlation. If AAPL stock moves drastically up or down, QCOM should follow based on this analysis. In the other direction, however, AAPL is far less dependent on QCOM in our analysis, showing that this correlation is generally one-sided. In other cases, high correlation factors between two companies might actually signal a negative correlation, where the increase in one company's stock will decrease a different one (GIVE EXAMPLE HERE).

4.2 Strengths of Approach

Our approach gives focused analysis of the correlation between two companies. Because we take specific keywords such as industry, total asset size, and location into account first, we get an factor that serves as a guide for the second part of the analysis. The MCA creates a symmetric matrix that is then funneled through the LDA. This first matrix from the MCA not only serves as a directional guide for the LDA, thereby increasing accuracy, it also increases overall speed. This initial MCA allows for us to have a fundamental economic intuition that pure content analysis would lack. Our LDA analysis then analyzes the remaining content of each company's Wikipedia pages, looking in particular for cross-company references, which establish and strengthen the directional correlations between companies. The LDA analysis not only captures diverse, unexpected correlatory factors that the MCA's factors may not capture, but also establish the asymmetry in pairwise directional relationships. The LDA then outputs our finalized asymmetric matrix, which gives directional relationships between each of the 500 companies. Since Wikipedia is user-based, more well-known companies that have more written about them and therefore have more data. Thus, the correlation between

larger companies is generally greater than it would be if the information had been more standardized. However, our approach takes care of this bias because of the first factor-based component that is independent of the amount of data included in their entire page. Moreover, larger companies will tend to correlate much more to each other as oppposed to smaller companies. This way, we keep this component without letting it overwhelm our data.

4.3 Weaknesses of Approach

Our main weakness lies in the fundamental characteristics of our data. Because we could only use Wikipedia pages, we missed out on a lot of important economic factors. Additionally, while it may be advantageous to weight certain factors and keywords differently depending on company sector, we do not have the appropriate research on each sector, and thus we refrained from making assumptions about small differences between market sectors.

4.4 Technical Performance

Our usage of Cython boosts speeds significantly due to the simplification of the programming language's architecture. While our two-part analysis takes longer than single approach methods of analysis, the increased accuracy in terms of usable correlation justify our increased run time. By limiting the number of factors we analyze to key, widely understood factors, we optimize our run-time performance.

5 Conclusions

By using factor and content-based approaches, we were able to produce a matrix we described above. In the future, we would explore diving into linked Wikipedia pages at the cost of more processing time, but additional accuracy.