Asset graph structural properties and parameter choices

Jacob Krol, CSCI 5352

05-05-2025

1 Introduction

1.1 Motivation

Asset graphs are networks linking financial assets, such as stocks, and are useful for discovering economic trends and anomalies. In this work, vertices are shares of companies called stocks and edges are unweighted connections to stocks with low or high correlation. Low correlation asset graphs have applications in portfolio diversification. Where, the goal is to minimize owning many stocks with correlated price to avoid losing all asset value all at once. Meanwhile, high correlation asset graphs have applications in pairs trading. Where, the goal is to find a pair of correlated stocks and wait to observe a divergence in price; one stock's price increases while the other decreases. Then, based on strong belief in correlation being restored, a stock trader can bet against the increasing stock and buy the decreasing stock for profit as their price trends later converge. This is pairs trading.

1.2 Related work

A pioneering study of asset graphs was Mantenga 1999 [2] which investigated stock clustering based on 1989-1995 SP 500 price data. Mantenga 1999 formalized stock pairwise comparison by computing daily changes in stock price, then using computing pairwise correlation of each stock price vector to build an asset graph. Their analysis focused on clustering the minimum spanning tree derived from the asset graph with applications in portfolio diversification.

Building upon Mantenga 1999, Onnela 2003 [3] studied the structural development of asset graphs by adding edges one at a time based on the highest ranking correlation stock pair. Onnela compared asset graphs to a random graph model which selected at random any stock pair during each edge addition step. Their results showed that the clustering coefficient is much higher in the early structural development of asset graphs compared to random graphs. Furthermore, Onnela used monthly price change vectors, compared to daily price change vectors used in Mantenga.

A similar random graph model was employed in Connor 2017 [1] to study the structure of microbial co-occurence networks. With count data of taxonomic abundances at different sampling sites, Spearman correlation was traditionally used to correlate a pair of sampling sites. However, this study showed that the choice of correlation threshold can lead to significantly different graph structure, such as before/after emergence of the giant component. And, a strong random graph model was used by shuffling the elements of underlying data matrix before correlation computation; this is in contrast to Onnela's shuffling of the correlation matrix. Also, Connor 2017 showed that adding fractional random noise to count data can break ties and affects the resulting microbial co-occurence networks.

A gap in the Mantenga and Onnela studies is that there was no investigation of how the time difference between price data affects the produced asset graph. It is an open question how different time windows would affect correlations and asset graphs.

Furthermore, Onnela's random graph model simply samples stocks at random and does not convey any randomness about the underlying price data. While, the Connor 2017 data matrix shuffling random process can test how random the underlying price data is.

1.3 Impact

This work studies the affect of choosing the time resolution of stock price data on asset graph structure. And, asset graphs are compared to a randomized data matrix graph model to investigate the randomness of underlying price data.

I found that time resolution noticeably affects asset graph structure. In particular, asset graphs derived from daily price changes are much less random than those derived from yearly price changes. I also contribute a perspective on low correlation asset graphs for direct applications in portfolio diversification. Finally, I quantified the tendency of stock correlation by company sector by measuring network modularity with respect to (w.r.t.) company sector node labels.

2 Methods

2.1 Stock data

Seven years of stock data spanning 2010-2016 of companies in the S&P 500 stock index were was collected from Kaggle (https://www.kaggle.com/datasets/dgawlik/nyse). 36 companies were discarded since they did not have price data for the full seven year period. The closing price (adjusted for splits) was used for all price calculations. And, the "GICS sector" was used to label a node's/company's sector in asset graphs.

2.2 Asset graphs

As in [REF], asset graphs link stocks based on correlation of their price differential data.

Formally, for each stock, the log difference in price after some number of days t yields a return vector, $\mathbf{r}^t \in \mathbb{R}^{\lfloor T-1/t \rfloor}$.

$$\mathbf{r}^t = ln(P_i(x)) - ln(P_i(x-t)) \tag{1}$$

Where, $P_i(x)$ is the price of stock i at time x.

Next, Pearson correlation of return vectors yields pairwise correlation of stocks i, j.

$$cor(i,j) = \frac{\langle \mathbf{r}_i^t \mathbf{r}_j^t \rangle - \langle \mathbf{r}_i^t \rangle \langle \mathbf{r}_j^t \rangle}{\sqrt{((\langle (\mathbf{r}_i^t)^2 \rangle - \langle \mathbf{r}_i^t \rangle^2)((\langle (\mathbf{r}_j^t)^2 \rangle - \langle \mathbf{r}_j^t \rangle^2)}} \tag{2}$$

Importantly, asset graphs are constructed by filtering the pairwise edge set by applying a correlation threshold to prioritize either high or low magnitude correlation asset subgraphs. However, correlation thresholds are arbitrary and can drastically affect network structure. To address this, asset graph structure is studied at a variable range edges, which is equivalent to letting the correlation threshold vary.

In more detail, asset graphs constructed by a growth procedure where nodes and edges are iteratively added by selecting the highest correlation stock pair at each step (Algorithm 1).

Given the computational complexity of computing structural statistics for many asset graphs in the experiments, the maximum number of edges, m, was upper bounded at $m <= 1000$. Empirically, major structural changes, such as the emergence of the giant component, usually occurred within $m \in [1, 1000]$. However, $1000/\binom{464}{2}$ is a very small fraction of the possible edge set, and a larger amount of m should be explored in future work.

Algorithm 1: Build asset graph

```
Input:
1. Empty graph $G = (V, E) : V = E = \emptyset$
2. Stock correlation matrix $X$
3. $m = 1000$
Output: Asset graph $G'$
for $i \leftarrow 1$ to $m$ do
    $i, j \leftarrow \arg\max_{i,j} X[i, j]$ only consider
     lower triangle
    $E \leftarrow E \cup \{(i, j)\}$
    $V \leftarrow V \cup \{i, j\}$
    remove pair $(i, j)$ from future consideration
end
$G' = (V, E)$
return $G'$
```

2.3 Time window parameter

t is a user-defined parameter quantifying the number of days between price differential calculation Eq. 1. Intuitively, a low t yields a high-dimensional, high-resolution return vector while high t yield low-dimensional, low-resolution return vector. By analogy, a lower t provides more data points similar to how a higher frame rate provides higher video quality. t is an important parameter. It modifies the resolution of stock time series data used in computing correlation and building asset graphs.

In the described experiments, "daily asset graphs" were defined by $t = 1$ while "yearly asset graphs" used $t = 240$; there are usually 240 trading days in a year. An important takeaway is that changing t also changes the underlying price data, correlation matrix, and resultant asset graph.

2.4 Low correlation asset graphs

"Low correlation asset graphs" are asset graphs where edges are added based on low magnitude correlation. To do this, the correlation matrix, $\mathbf{X}$, was transformed element-wise by $f(\mathbf{X}_{ij}) = 1 - |\mathbf{X}_{ij}|$. This transformation coerces the correlation values $\in [-1, 1]$ to a $[0, 1]$ range. Original correlation values close to 0 become close to 1 in the transformed matrix. Therefore, low correlation asset graphs were built by applying Algorithm 1 to $\mathbf{X}_{transformed}$.

2.5 Random graphs

Random graph models were used to observe whether asset graph structure was different than expected by random chance. To build a random graph, the underlying stock price data matrix, $\mathbf{P} \in \mathbb{R}^{stocks \text{ x } \lfloor T-1/t \rfloor}$, elements was shuffled 1000 times using Fisher-Yates shuffling (NumPy implementation). Then, pairwise Pearson correlation (Eq. 2) was computed using the shuffled data, and the build algorithm was applied (Algorithm 1).

2.6 Company sector modularity

Modularity was used to measure the tendency of stocks in the same sector to be linked. Each stock vertex was labeled with a categorical variable indicating their sector membership. Using the company labels, modularity was computed as

$$Q = \frac{1}{2m} \sum_{i,j} \left[A_{ij} - \frac{k_i k_j}{2m} \right] \delta(c_i, c_j) \tag{3}$$

where k_i and c_i was the degree and sector label of vertex i, respectively.

A modularity of 0 indicates sector assortativity is no different than expected at random. While, a modularity of 1 indicates high assortativity by sector, and -1 indicates high disassortativity by sector.

In network visualizations, nodes are colored by their sector using the following key. Consumer Discretionary: lightred, Consumer Staples: lightblue, Energy: green, Financials: orange, Health Care: cyan, Industrials: magenta, Information Technology: yellow, Materials: brown, Real Estate: pink, Telecommunications Services: teal, Utilities: gray.

2.7 Structural statistics

Global graph statistics measured the structure of asset and random graphs at various t and m. Clustering coefficient (C), mean degree ($\langle k \rangle$), size of largest connected component ($|LCC|$), and the number of connected components (num. components) were all computed using standard formulae. Diameter (l_{max}) however, was computed within only the LCC.

3 Results

3.1 High correlation asset graphs

hypothesized that asset graph structure was sensitive to the time resolution of price differential vectors used to compute correlation. To test this hypothesis, an experiment was conducted to measure global structure statistics (clustering coefficient, number of connected components, etc.) as the network grew by adding an edge between two stocks with the highest correlation. The expected result was that structure will be noticeably different for asset graphs with different time resolution. Also, it was expected that asset graph structure will be noticeably different than their corresponding random graph model.

Figure 1 shows that there exists noticable differences between high correlation daily asset graphs and yearly asset graphs; also, daily asset graphs appear less random.

Daily asset graphs, compared to yearly asset graphs, had a much higher clustering coefficient (C), modularity w.r.t company sector, and many more connected components at high m (Figure 1, columns 1 and 2). A snapshot of daily asset graph at $m = 100$ shows many three-way correlations particularly in the finance (orange nodes) and real estate (pink nodes) (Figure 1, column 1, row 1).

Similarly, daily asset graphs differ from their random graph model with much higher C, modularity, and number of components. In the random daily graph, a giant component emerged at $m \approx 800$, but the asset graph maintained a large number of components beyond $m > 800$. This was illustrated by the log_2 number of components in Figure 1 (columns 1 and 3, row 3, red line).

Meanwhile, the yearly asset graph was structurally similar to its random graph with $\forall_m > 200 : C \approx 0.4$ and low modularity (Figure 1, columns 2 and 4). However, the yearly asset graph still maintained more connected components at high m, compared to its random counterpart (Figure 1, columns 2 and 4, row 3, red line).

In summary, high correlation asset graphs were structurally sensitive to the time resolution of price differential data. Daily asset graph structure was less random compared to yearly asset graphs. Particularly, daily asset graphs had high modularity; meaning, edges commonly occurred between companies in the same sector. A key difference between high correlation asset and random graphs was that the giant component emerged much earlier in random graphs.

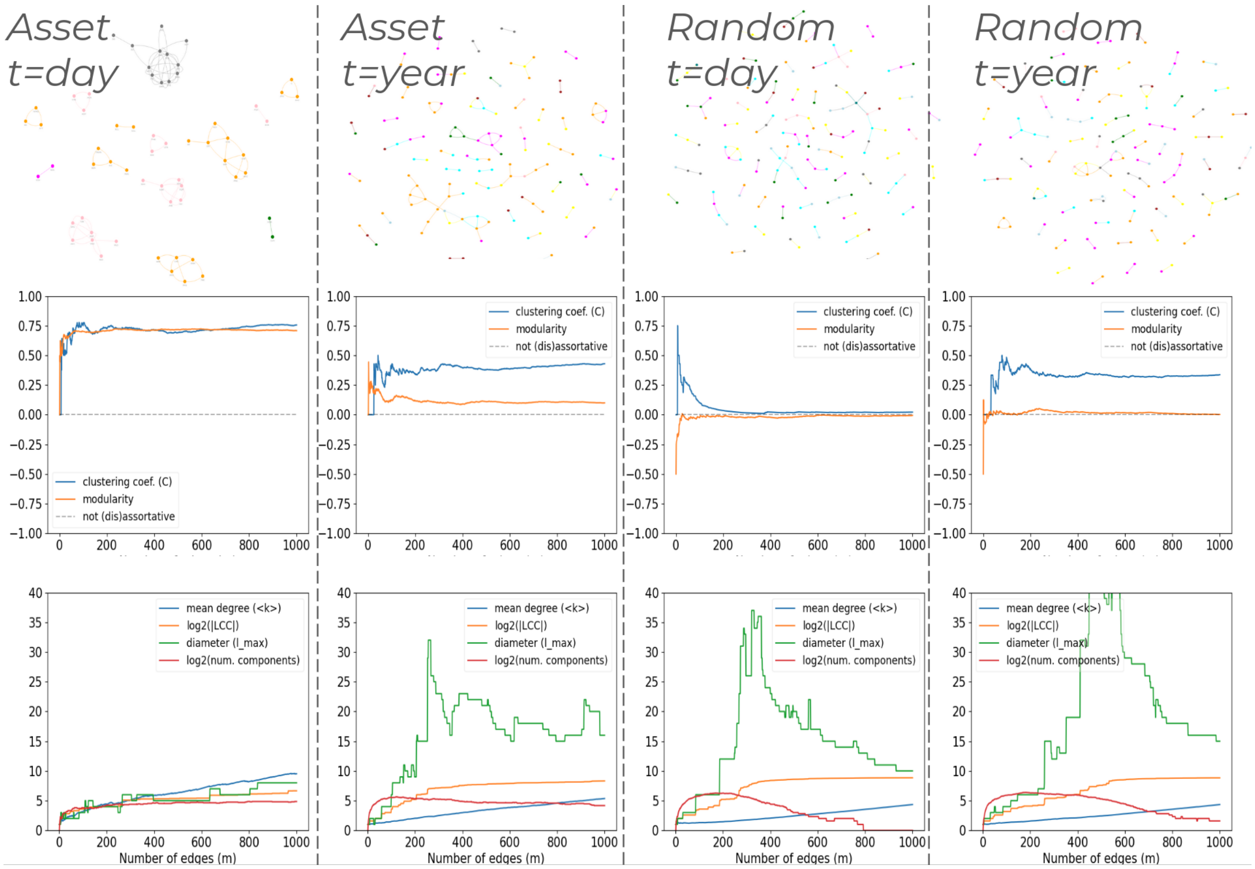

Figure 1: **High correlation asset and random graph structure as a function of edges** (m). Edges are added by selecting the *highest*, non-self correlation in the correlation matrix. For 2D line plots, the horizontal axis is the number of edges in the graph. **Columns** separate asset and random graphs by their time resolution (daily or yearly price difference). **Row one** is a network snapshot at $m = 100$. **Row two** shows the clustering coefficient (blue line), modularity (orange line), and baseline modularity (dashed line). **Row three** shows the mean degree (blue line), log_2 size of largest connected comoponent (LCC, orange line), diameter w.r.t. the LCC (green line), and log_2 number of connected components (red line).

3.2 Low correlation asset graphs

With applications in portfolio diversification, this experiment measures how time resolution of price differential vectors affects the structure of low magnitude correlation asset graphs (see Methods). Structural statistics of asset and random graphs were recorded as the network grew by adding an edge between two stocks with the lowest magnitude correlation. Again, the anticipated outcome was that asset graph structure would be non-random and depend on the time resolution of price data. From this experiment, I found that daily asset graphs had noticeably different structure compared to random graphs. And, yearly asset graphs were very similar to random graphs.

Notably, daily asset graphs had negative modularity while yearly asset graphs and random graphs had ≈ 0 modularity (Figure 2, row 2, orange line). The negative modularity contrasts the high correlation daily asset graph; meaning, low correlation asset graphs tended to have edges among stocks in different sectors. Also, the giant component of daily asset graphs emerges much earlier ($m \approx 250$) compared to yearly asset and random graphs. This was shown by the low number of components in daily asset graphs compared to others (Figure 2, row 3, red line).

Interestingly, the daily asset graph snapshot visualization (Figure 2, column 1, row 1, brown central node) showcases that Newmont Corporation (a gold/rare-metal producer) has low correlation to many companies. This supports the common ideology that rare metals are a useful hedge against other common assets.

In contrast, yearly asset graphs are structurally similar to their random graph counterpart by all metrics (Figure 2, columns 1-3). The yearly asset and random graphs all had modularity and C ≈ 0 (Figure 2, columns 1-3, row 2), and the giant component emerges much later (Figure 2, columns 1-3, row 3, red line).

In summary, low correlation asset graphs were indeed sensitive to the time resolution of their price data. Where, yearly asset graphs were close to random while daily asset graphs had distinct properties. Daily asset graphs were slightly disassortative by sector and well-connected as the correlation threshold was loosened.

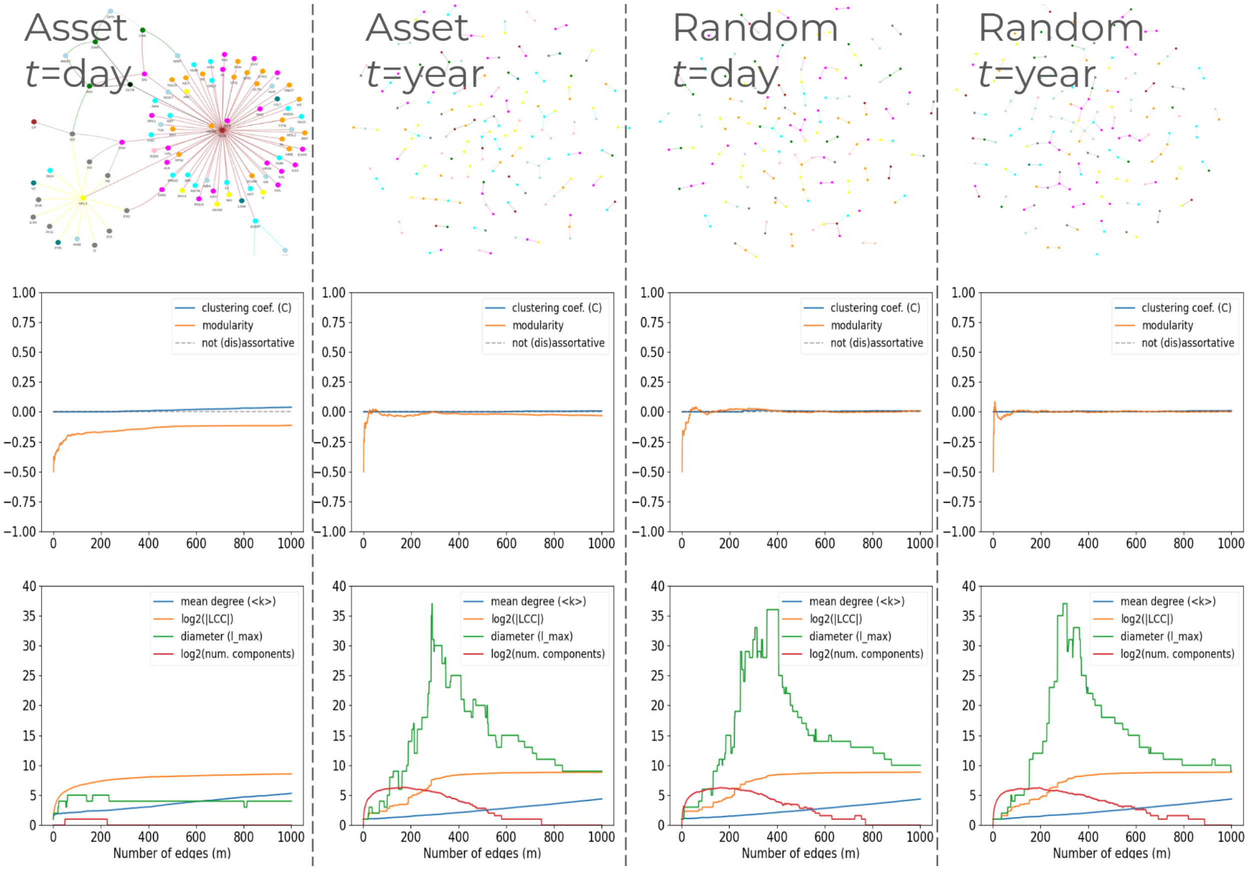

Figure 2: **Low correlation asset and random graph structure as a function of edges** (m). Edges are added by selecting the *lowest* magnitude, non-self correlation in the correlation matrix For 2D line plots, the horizontal axis is the number of edges in the graph. **Columns** separate asset and random graphs by their time resolution (daily or yearly price difference). **Row one** is a network snapshot at $m = 100$. **Row two** shows the clustering coefficient (blue line), modularity (orange line), and baseline modularity (dashed line). **Row three** shows the mean degree (blue line), log_2 size of largest connected comoponent (LCC, orange line), diameter w.r.t. the LCC (green line), and log_2 number of connected components (red line).

4 Discussion

The experimental results supported the hypothesis that time resolution, t, of price differential data affects asset graph structure. High correlation daily asset graphs were assortative by sector, and the clustering coefficient was high early in the graph development compared to random. Meanwhile, in low correlation daily asset graphs the giant component emerged much faster than random, and stocks were disassortative by sector. Additionally, a gold company (Newmont Corporation) was a very central node in the low correlation daily asset graph at $m = 100$. In contrast, yearly asset graphs were structurally similar to random. I hypothesize that yearly asset graph randomness is due to the lower dimensionality of the price differential vectors. Since, the dimensionality of price differential vectors is inversely proportional to t.

The non-randomness of daily asset graphs suggest that asset graphs are more useful for short-term strategies rather than long-term strategies. Yearly asset graphs would be ideal for long-term investment strategies, yet they are structurally similar to a random data generating process which makes them less predictable. Meanwhile, daily asset graphs have characteristic and non-random structure in both high and low correlation asset graphs.

Future work should 1) formalize the random graph models into null distributions of structural statistics, 2) explore a larger range of m, 3) use high time resolution asset graphs to quantify risk in real portfolios and simulate pairs trading performance. Null models of structural stats can be produced by repeating the matrix shuffling over a large number of permutations to generate a null distribution for each statistic. A larger range of m can easily be explored with more compute time and parallelization of stat computation with different asset graph parameters (e.g., m, t). Historical portfolios could be represented as simple graphs and compared to correlation based asset graphs. Finally, the profitability of pairs trading using high correlation daily asset graphs information could be simulated by simply applying it to real-time stock data.

5 References

References

[1] Neal Connor, Albert Barberán, and Aaron Clauset. Using null models to infer microbial cooccurrence networks. *PLoS ONE*, 12(5):e0176751, 2017.

[2] R. N. Mantegna. Hierarchical structure in financial markets. *European Physical Journal B*, 11:193–197, 1999.

[3] J.-P. Onnela, A. Chakraborti, K. Kaski, J. Kertész, and A. Kanto. Asset trees and asset graphs in financial markets. *Physica Scripta*, 2003(T106):48, 2003.

6 Appendix

Code located at

`https://github.com/jakekrol/csci-5352-network_analysis_and_modeling/tree/main/proj`