Beyond correlation networks

Nicholas Huang, Leo Keselman, Vincent Sitzmann
{sitzmann, lkeselman, nykh}@stanford.edu

1. Introduction and Problem Definition

Stock correlation networks are a widely used model to analyze stock markets. However, using only the stock correlation networks to make assertions about a market is naïve because they only capture the result of historic correlations, and not their underlying dynamics. We propose using a collections of other network topologies to obtain a diverse set of feature-rich networks that have much more potential for studying market behavior. Examples for such topologies are networks based on business objective similarity, market capitalization or geographic location.

Having assembled such a variety of topologies, we benchmarked their performance against classical correlation networks on a number of prediction tasks, such as link prediction and anomaly detection. This allowed us to verify that the assembled networks capture salient information of the underlying stock market dynamics. Additionally, we evaluated how standard graph algorithms for link prediction and graph visualization perform on the new topologies.

Additionally, we perform historical analysis using stock market data from the 1920s through 2015, comparing the topologies and structures of correlation networks through time. To the best of our knowledge, this is the first work looking at how basic properties of stock correlation networks change over long periods of time. Contrary to typically scale-free networks, which grow denser over time, stock correlation networks over the 20th century exhibit a large growth in node number.

2. Related work

2.1. Stock correlation networks

The first approach to capture the structure of the stock exchange in a graph was established by [9], who proposed an undirected "stock market correlation network".

In this definition, nodes represent equities and the edges between nodes are determined by calculating pairwise correlation of stock prices within a certain time and then either thresholding the absolute correlation with a threshold value $\theta \in [0, 1]$ or applying a graph-building algorithm such as the minimum spanning tree method.

Many improvements on this method has been proposed, such as picking correlation metrics that are robust to general trends in the market [14] or ways to captures correlations with lagged response [15].

[13] proposed extending the pure price correlation to a cross-correlation of price and volume. Studies have also revealed a number of interesting properties.

[3] discovered the node distribution in stock correlation networks follows a power-law for sufficiently high threshold θ. They also discovered the independent sets in the stock network to be generally small, which indicates it is difficult to design completely diversified portfolios.

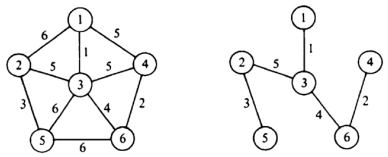


Figure 1: Problem of using a minimum spanning tree to build the stock correlation network: The edge between securities 5 and 3 with weight 6 is lost although it has only a slightly higher weight (=lower correlation) than the edge between securities 2 and 3. This greatly alters the true underlying topology of the graph, since, for instance, price predictions of security 3 are now not dependent on security 5 anymore. Image obtained from [1], chapter 7

All the aforementioned work is similar in that it only considers metrics that are observed on the stock market itself to build the financial network. This purely correlation-based approach, however, cannot explain causal relationships between the traded securities. This is a great obstacle against applications such as portfolio building, since the model lacks insight as to why stocks are (un-)correlated in the past and if these conditions still hold true. Further, much work in this field builds these networks using minimum spanning trees. This is problematic, since the minimum spanning tree is not guaranteed to preserve the true topology of the graph, as Figure 1 demonstrates.

2.2. Link Prediction and Expanding Network Dynamics

Link prediction is an area of research in social and information graphs that studies how edges evolve over time. In social networks, link prediction is used in recommendation system, but it has also been applied to stock correlation graphs.

[8] was one of the seminal papers in the field of link prediction. It presented a suite of methods, ranging from simple Jaccard similarity metrics to variations on PageRank. They put emphasis on the robustness requirement for good link predictors are capable of being robust to a few spurious edges (for example, authors collaborating across disjoint fields in their author dataset).

[16] presented a method for nonparametric link prediction. Specifically, they evaluated their method on stock market correlation graphs for the S&P500 stocks (500 high-performing stocks). However, for stock market data, their method wasn't as strong as that of using Katz measures [8]. This may be due to methodological issues, as they only evaluate large-capitalization stocks, which, due to their size, don't exhibit much evolution over time.

Other methods looked at a network graph as a continually evolving system, where links can change as time flows. [6] used a probabilistic model where the graph is broken into subgraphs, and then pairwise models for prediction are built explicitly for each pair in each subgraph. This approach scales well, as it takes the $O(n^3)$ modeling problem of all pairwise interactions and turns it into a $O(kn^2)$ model of distinct subgraphs. However, this partitioning approach forces explicit group formation which is unable to model nodes belonging to multiple groups simultaneously.

[2] presented a predictor that predicts not only a connection but also its strength. It formulates the link prediction problem as a supervised training of a random walk algorithm. The supervision is used to force high PageRank scores for nodes to be connected, and lower score otherwise. This analytical formulation was interesting, as it can easily be extended to include temporal regularization for a historical network. This approach is fast, learns a small set of interpretable weights, and might be very effective for understanding temporal dynamics in financial networks.

2.3. Anomaly Detection and feature representation

Anomaly detection is the study of finding anomalous element in data. In the context of network study, this could mean the anomaly within a network, for example node with high degree, but in our context it refers to the anomalous behavior of a network. Chandola et al. provided a general survey of anomaly detection. We adopted basic idea about anomaly detection in finding a feature representation for the network and use various classifier to detect anomaly. However, we could not find more data on vector representation of a network. This could be because network is itself a structure of great dimension, and representing it as a vector was not a very common practice. Therefore, we decide to use aggregate measurement of network as representation.

2.4. Kronecker graphs for graph topology approximation

[7] described Kronecker graph and stochastic Kronecker graph model, as well as an efficient algorithm to fit a Kronecker graph to a given network. `KronFit` is an implementation of the Kronecker graph fitting algorithm developed using SNAP.

2.5. Semantic Similarity

In order to capture business interest, shareholder overlap, or other such semantic overlap in a graph topology, we need some scoring function. Recent research in natural language processing has demonstrated the efficacy of word-vector embeddings [12] in compactly capturing semantic distance. These GLoVe vectors have been shown to capture linear subspaces between companies and their CEOs, based on a model learned from Wikipedia. [4] built a semantic-similarity graph with such embedded word vectors to perform analysis of semantic structures; we hope to utilize them to capture relationships between businesses.

3. Methods and Algorithms

3.1. Data Collection

For the stock price data, we used the Center for Research in Security Prices (CRSP) dataset, obtained from the Wharton Research Data Services (WRDS), University of Pennsylvania. This dataset comprises of daily prices for over 26,500 stocks listed on the New York Stock Exchange (starting in 1925), the New York Stock Exchange Archipelago Exchange (starting in 1962), the American Stock Exchange (starting March 2006) as well as the NASDAQ (starting in 1972). It is high-quality and professionally curated. To get the geographical information about a company, we used the CRSP-Compustat Merged Dataset published by the same source, which links the daily prices of CRSP dataset with additional information about the companies.

For any time period, we can construct a correlation network by calculating the cross-correlation of daily average return and take a threshold on the absolute value of correlation. Because stock correlations may change over time, we need to define the time period over which to construct the correlation network. Following the convention of previous study we calculate the correlation ρ_{ij} between time series i and j as

$$\rho_{ij} = \frac{\sum_t [(x_i(t) - \bar{x}_i)(x_j(t) - \bar{x}_j)]}{\sqrt{\sum_t (x_i(t) - \bar{x}_i)^2} \sqrt{\sum_t (x_j(t) - \bar{x}_j)^2}} \quad (1)$$

where $\bar{x}_i$ represents the mean value of discrete time series x_i. We then construct a network where node i and j is connected if the absolute value of ρ_{ij} is greater than a threshold.

For this project, we have largely followed the convention set up by previous work in constructing a correlation network, with the modification that we use the absolute value of correlation as the criterion for connection. Future work could look at alternative measures of correlation such as time-lag correlation or nonlinear correlations.

To build a network as a reference of study, we selected the daily average return data from the full year of 2014 of New York Stock Exchange (NYSE) and NASDAQ trading data, including only stocks that have valid and varying data in the time period. Figure-2a shows the degree distribution of this network, which exhibits a clear power law distribution. We call this network *the canonical network* in subsequent sections and use it as the typical representation of correlation networks.

3.2. Alternative Networks

Our goal is to explain connections in the correlation network with alternative information not directly related to the price. This would allows us to build a powerful, independent predictor. One potentially useful source of information was geography. Different countries and states have different economic and regulatory policies as well as clustering effect that can affect the performance of companies based mainly in them. We extracted geographical information from the CRSP-Compustat dataset and defined a network structure where two stocks are connected if they share a state. This definition creates a network that is a collections of clustering, one for each state or country.

Another piece of information we gathered was the semantic information derived from the company name. We used the *GloVe* distributed representation of English words as the basis for the semantic feature vector [12]. The pretrained GloVe vectors available from the original source was trained by crawling Wikipedia pages, which we consider to encode useful information about companies. Intuitively, if two companies often appear together in a Wikipedia article or web page, there is a higher possibility that they conduct related business and their stock price might thus be more correlated. We applied a ranking selection method to select only those pairs of company names with closest distance in the GloVe space. This allowed us fine control on the number of edges we want the graph to contain. As we can see in Table-1, we chose to have the semantic network match canonical network in dimension as close as possible. The resulting semantic network exhibits power law distribution and also shows similar properties as the canonical network. As Figure-2c shows, the semantic network also have similar degree distribution with the canonical network.

We also were interested in constructing graph weights with alternative properties than simple correlation. To do this, we incorporated the comparative size of companies in the stock market by creating what we call the *Market capitalization network*. We processed the canonical network into a directed version, where each edge in the canonical network becomes a directed edge, pointing from the company with greater total market capitalization to the one with smaller capitalization. As shown in Table-1, it has identical properties to the canonical network except diameter.

3.2.1 Kronecker Graph

Another useful way of understanding a network is to fit a Kronecker Graph model to it. We used `KronFit` algorithm to efficiently fit a Kronecker graph to each of the networks mentioned earlier [7]. Table-2 shows result of fit, where the fitting result is a parameter matrix

$$\Theta = \begin{pmatrix} \alpha & \beta_1 \\ \beta_2 & c \end{pmatrix}$$

By the Stochastic Kronecker Graph theorem, we see none of the networks thus introduced is connected (geography network is by definition a collection of separate clusters). We can validate this by checking the largest weakly connected component in the canonical graph, which only represents about one third of the nodes in the network.

On the other hand, the semantic network exhibits the "giant component" shape that is typical of a real network. This justifies our construction of the alternative networks.

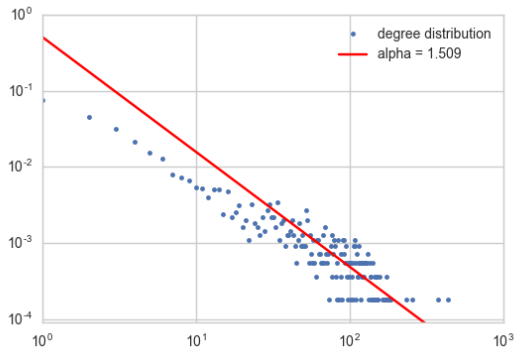
To evaluate how well these information explain salient information in the correlation network, we incorporated them in tasks like link prediction and anomaly detection, as discussed below.

3.3. Link Prediction

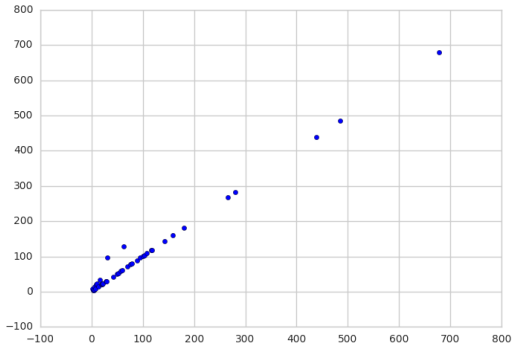
First we evaluated link-prediction performance. It can reveal either the robustness of a topology (when evaluating a given time instance of a graph), or serve as a method of predicting future behaviors (when compared over pairs of consecutive time instances). The former can be applied to any graph, while the latter requires correlation-based graphs (as geographical graphs and the like tend not to change over time). We performed multiple classes of link prediction methods and as described below.

3.3.1 Neighborhood Techniques

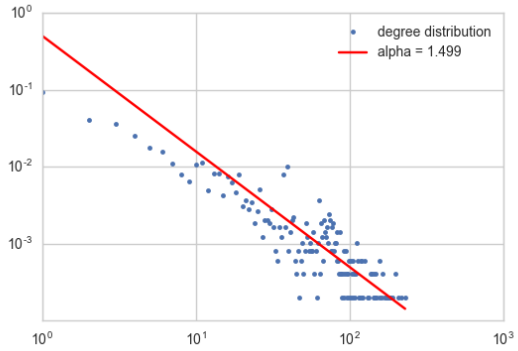
It has been shown that triangle closing is a valid measure for collaboration networks, where x, if x and y share many common neighbors at time t, they have good odds of collaborating together at some time after t [10]. From this line of work, there are multiple methods that look at common neighbors and possible triads as ways of predicting links.



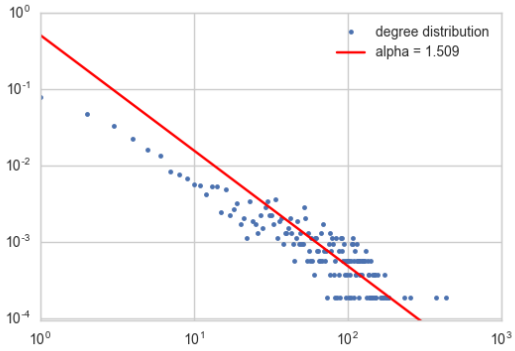
(a) Canonical Network



(b) Geography Network



(c) Semantic Network



(d) Market Cap Network

Figure 2: Degree distributions of different network structures.

Network	Nodes	Edges	Power Law Coefficient	Diameter	Average Cluster Coefficient
Canonical	5483	23416	1.499	5.80	0.19
Geography	4370	620608	–	0.90	1.0
Semantic	4892	23416	1.509	5.69	0.23
Market Cap	5483	23416	1.499	6.83	0.19

Table 1: Properties of different network structures

Network	α	β_1	β_2	γ
Canonical	1	0.531	0.531	0.215
Semantic	1	0.544	0.518	0.233
Geography	1	0.601	0.601	0.560

Table 2: Resulting parameters of fitting Kronecker graph

Common Neighbors Probably the simplest indicator to score whether two nodes should share an edge is to count the number of shared neighbors between the two nodes as, with $\Gamma(x)$ indicating neighbors of x

$$C(x, y) = |\Gamma(x) \cap \Gamma(y)| \tag{2}$$

Common Neighbors with Weights A modification of the common neighbors indicator is a weighted common neighbors indicator, where nodes with larger weights exert more influence. We formulate this as

$$C_w(x, y) = |\Gamma(w_x x) \cap \Gamma(w_y y)| \tag{3}$$

Jaccard similarity The Jaccard similarity intuitively captures the ratio of neighbors that are shared by two nodes to the number of total neighbors of those nodes:

$$J(x, y) = \frac{|\Gamma(x) \cap \Gamma(y)|}{|\Gamma(x) \cup \Gamma(y)|} \tag{4}$$

Method	Graph	Accuracy	Energy	Energy/Random
Common	Canonical	0.27	0.93	4.11
CommonW	Canonical	0.34	0.94	4.16
Jaccard	Canonical	0.00	0.72	3.18
Common	MarketCap	0.28	0.94	4.16
CommonW	MarketCap	0.34	0.95	4.20
Jaccard	MarketCap	0.00	0.73	3.23
Common	Semantic	0.21	1.11	1.00
CommonW	Semantic	0.08	2.81	2.52
Jaccard	Semantic	0.00	1.44	1.29
Common	Geographic	1.00	1.00	1.02
CommonW	Geographic	1.00	1.00	1.02
Jaccard	Geographic	1.00	1.00	1.02

Table 3: Results from various link prediction methods and how well they perform at predicting graph structure and predicting correlation

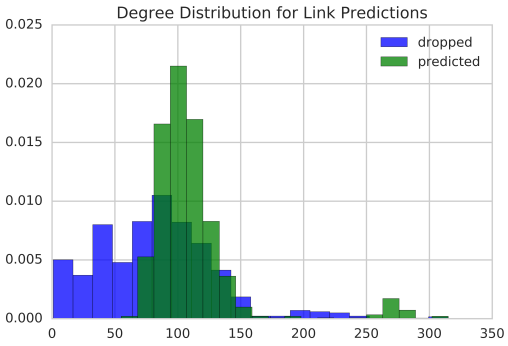


Figure 3: Degree Distribution for Link Prediction. Blue region corresponds to degrees of dropped nodes. Green region corresponds to predicted nodes.

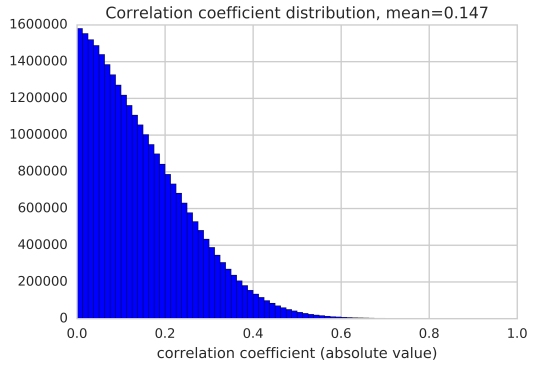


Figure 4: Correlation Coefficient Distribution

This scoring method penalizes nodes connections between nodes which have high degree already, and instead treats each link prediction based on the size of it's shared neighbor set.

3.3.2 Path Techniques

Another technique for computing similarity is based on looking at the paths between nodes and various metrics on top of the distribution of path lengths. As before, the simplest link predictor is simply to score nodes based on their shortest path in the network.

We implemented these techniques, but they tended to perform very poorly. In our testing method, described below, the results were no better than random. Alternatives to simple shortest path were also unsucessful, as they tend to (such as with the Katz [5] measure) simply damp further connections.

3.4. Anomaly Detection

Anomaly detection is the problem of finding anomalous dynamics or properties of the network. In this project we would like to use network structure features to predict stocks that are high-performing stocks and analyze whether features from of one year (the *reference network*) can imply performance of that stock in the next year. In this sense, we define a *high-performer* as a stock whose price rises by at least 10% above the average price increase of the overall stock market. As an example, if the average stock market price increases by 10%, a high-performing stock must increase by at least 20% in price.

We posed this problem as a binary classification problem. In the first step, we extracted features of a stock in the reference network. We extracted the following node-specific features:

1. PageRank [11]
2. Node degree

3. Betweenness centrality
4. Farness centrality
5. Closeness centrality
6. Degree centrality

We then extracted the same features for both the syntactic and the geographical networks. We further considered the following domain-specific features:

1. Split-adjusted market capitalization
2. Split-adjusted market price
3. Which exchange (New York Stock Exchange vs. Nasdaq)
4. Shares outstanding

We then fed these features into a binary classifier. We considered linear Support Vector Machine classifier (SVM), logistic regression, as well as Random Forests (of 100 trees with maximum depth 8) as classifiers. We trained and tested our classifier on the stock correlation networks from the years 1925 to 1999. After that, we tested the effect of including features from previous years to study the effect of time-dependent knowledge in anomaly detection.

3.5. Historic graph similarity

We analyzed how a stock correlation network compares to historical stock correlation networks in the past. We hope this will be useful for identifying possible boom periods and recessions. We implemented this by extracting a number of graph-specific features, such as network diameter, the average cluster coefficient, the node and edge count, the average and median degree, as well as the power law coefficient α resulting from a power-law fit. These features are standardized by subtracting the mean and dividing by the standard deviation. We then calculated the euclidean similarity of these feature vectors across years.

As additional feature to consider, we fit Kronecker graphs to each year's network. This yields the Kronecker base matrix, as described in 3.2.1. This not only yields another distance metrics between stock correlation networks of different years, namely the $l2$ distance of the 2 Kronecker parameter matrices. It also allows us to investigate the development of the kronecker fit parameters α, β_1, β_2 and γ over time, which encode valuable information about graph structure.

4. Results and Findings

In order to evaluate the quality of various network topologies, we setup various experiments with numerical benchmarks.

4.1. Link Prediction

In order to test the strength of various topologies, in terms of both their own robustness and in terms of predicting stock correlation, we implement Link Prediction.

Our experimental setup is fairly straightforward. We take a graph, drop one percent of its edges and run a link prediction algorithm. Algorithms are scored on two metrics, how many of the dropped edges they predicted successfully and how much correlation energy the predicted edges contribute. If the network is a stock correlation network than the latter metric can only be optimized by successful prediction, but in other topologies, the link prediction algorithm may pick nodes that have higher correlation energy than those randomly dropped. We repeat these experiments 20 times and report their results in Figure-3.

On the correlation networks, the common neighbors predictor predicted about 30% of links correctly, with a slightly higher score for its weighted alternative (as described earlier). Of the mispredicted edges, many of them still had high correlation energy, and we were able to capture more than 90 % of the correlation energy that was lost. Results were slightly better with weights, and improved with our market capitalization.

However, examining the predictions of these networks, as shown in figure 3, they tended to have very high degree.

To cope with this issue, we tested Jaccard similarity metric, which only looks at percentage of shared common neighbors, not absolute number. However, this method predicted none of the dropped edges, although it was capable of recovering over 70% of the lost correlation network. Of interest, it often predicted nodes between large, well known companies, such as Apple and Berkshire Hathaway.

On correlation networks, all of these methods were capable of predicting edges with strength 3 to 4 times that of an average possible edge in the network (the distribution of which can be seen in figure Figure-4.

With semantic networks, we saw behavior that was very surprising. Common neighbors predicted twenty percent of dropped edges, and the predicted edges tended to have strong correlation. The weighted neighbor predictor that we presented got worse accuracy, but was capable of recovering a large fraction of the stock correlation energy (two and half times that of a random performing link predictor). This shows that semantic networks capture some information that is present in correlation networks, without needing to refer to correlations themselves.

On geographical networks, due to our clustering construction, the link predictors succeed in recovering all of the dropped edges.

4.2. Anomaly Detection

Our dataset comprised 74 stock correlation networks covering years 1925 to 1999. This translates to a total

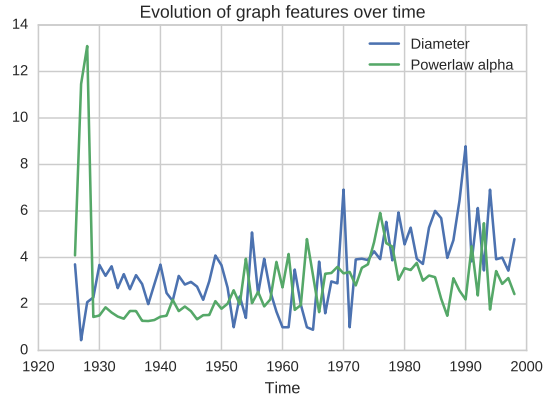


Figure 5: Showing how diameter and best-fit power law parameter (α) evolve over time

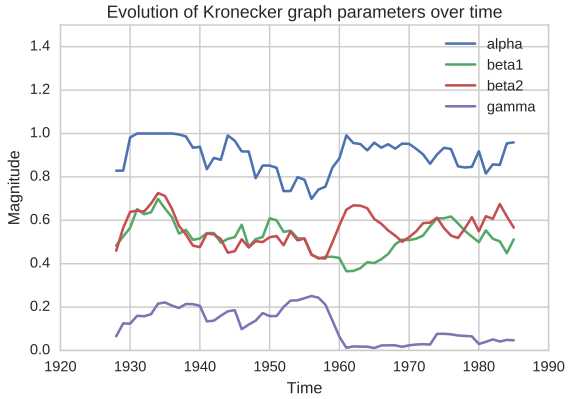


Figure 6: A view of stock markets evolve over time as parameterized by best-fit Kronecker Graph models

of approximately $206,000$ stocks represented by nodes, of which approximately 10% each year were identified as high-performers. A classifier assigning the high-performer label at random would thus achieve an accuracy of around 90%.

In our first experiment, we trained different classifiers to predict next year's high performers based on this year's stock correlation network graph features concatenated with the graph features of the geographical and semantic networks. We found that none of our classifiers outperformed the random baseline.

Next, we test the performance of classifiers trained on both the graph features and the domain-specific features. We thus concatenate those two feature vectors, yielding a feature vector of a length of 10. We again train the same classifiers as above, but now also include a SVM with a radial basis function kernel to allow for the fitting of a potentially non-linear decision boundary. However, the results remain the same - none of the classifiers outperformed the random baseline.

Lastly, we concatenated feature vectors of the previous 5 years to allow our classifier to consider temporal information. Again, however, none of the classifiers were able to outperform the random baseline.

These results are intuitive - while it is an ongoing discussion if markets are completely efficient, they can certainly be assumed to be efficient enough to price in all information that could be extracted by a linear classifier. It is thus likely impossible to robustly predict high performers in the market with high confidence.

4.3. Historic graph similarity

In figure 5, we plot the development of two remarkable measures for this experiment: The diameter of the networks as well as the α of the power law fit. The development of the diameter is remarkable, because in real-world datasets, networks usually grow denser and the diameter increases - this does not seem to be the case for stock correlation networks. We further compute the Kronecker fits for each graph in the timespan from 1925 to 1999. 6 depicts the development of the Kronecker graph parameters over time. We find that $\beta+\gamma < 1$, and thus that stock correlation networks are never connected over this time period. Further, we find that since α is always close to one, $(\alpha+\beta)*(\beta+\gamma) > 1$ and thus, although the correlation threshold is held constant over time, the stock correlation networks keep natural properties.

5. Conclusion

In this project, we have dissected the classic stock correlation network as a way to analyze stock markets. By developing geographic and semantic networks and using them to predict links in the stock correlation network, we have shown that feature network can explain some correlations in the stock correlation network. We have analyzed the use case of stock correlation networks for predicting high-performers in stock markets and have shown that stock correlation networks themselves are unlikely to contain the information necessary for this feat. We used the Kronecker graph as a meaningful way to describe developments of networks over time and have used this analysis tool to depict the development of the stock market within the last century.

References

[1] A. V. Aho, J. E. Hopcroft, and J. D. Ullman. *Data Structures and Algorithms*. Pearson, 1st edition, 1983.

[2] L. Backstrom and J. Leskovec. Supervised random walks: predicting and recommending links in social networks. In *Proceedings of the fourth ACM international conference on Web search and data mining*, pages 635–644. ACM, 2011.

[3] V. Boginski, S. Butenko, and P. M. Pardalos. Statistical analysis of financial networks. *Computational statistics & data analysis*, 48(2):431–443, 2005.

[4] M. Faruqui, J. Dodge, S. K. Jauhar, C. Dyer, E. Hovy, and N. A. Smith. Retrofitting word vectors to semantic lexicons. *arXiv preprint arXiv:1411.4166*, 2014.

[5] L. Katz. A new status index derived from sociometric analysis. *Psychometrika*, 18(1):39–43, 1953.

[6] M. Lahiri and T. Y. Berger-Wolf. Structure prediction in temporal networks using frequent subgraphs. In *Computational Intelligence and Data Mining, 2007. CIDM 2007. IEEE Symposium on*, pages 35–42. IEEE, 2007.

[7] J. Leskovec, D. Chakrabarti, J. Kleinberg, C. Faloutsos, and Z. Ghahramani. Kronecker graphs: An approach to modeling networks. *Journal of Machine Learning Research*, 11(Feb):985–1042, 2010.

[8] D. Liben-Nowell and J. Kleinberg. The link-prediction problem for social networks. *Journal of the American society for information science and technology*, 58(7):1019–1031, 2007.

[9] R. N. Mantegna. Hierarchical structure in financial markets. *The European Physical Journal B-Condensed Matter and Complex Systems*, 11(1):193–197, 1999.

[10] M. E. J. Newman. Clustering and preferential attachment in growing networks. *Phys. Rev. E*, 64:025102, Jul 2001.

[11] L. Page, S. Brin, R. Motwani, and T. Winograd. The pagerank citation ranking: bringing order to the web. 1999.

[12] J. Pennington, R. Socher, and C. D. Manning. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, 2014.

[13] B. Podobnik, D. Horvatic, A. M. Petersen, and H. E. Stanley. Cross-correlations between volume change and price change. *Proceedings of the National Academy of Sciences*, 106(52):22079–22084, 2009.

[14] B. Podobnik and H. E. Stanley. Detrended cross-correlation analysis: a new method for analyzing two nonstationary time series. *Physical review letters*, 100(8):084102, 2008.

[15] B. Podobnik, D. Wang, D. Horvatic, I. Grosse, and H. E. Stanley. Time-lag cross-correlations in collective phenomena. *EPL (Europhysics Letters)*, 90(6):68001, 2010.

[16] P. Sarkar, D. Chakrabarti, and M. Jordan. Nonparametric link prediction in dynamic networks. *arXiv preprint arXiv:1206.6394*, 2012.