UNIVERSITY OF FLORIDA
UF JUR
JOURNAL OF UNDERGRADUATE RESEARCH ISSN:2638-0668

Network Approaches to Analyze the Dynamics of Financial Markets

Miranda Rose Lochner
College of Engineering, University of Florida

Faculty mentor: Panagote Pardalos, Department of Industrial and Systems Engineering

Abstract

Analyzing financial markets requires gathering large amounts of data and determining appropriate methods so that accurate and appropriate conclusions can be drawn. The purpose of this paper is to investigate network approaches to understand large amounts of financial data and the implications of different approaches. Creating a market graph has been used to analyze financial instruments, and prices fluctuations of stocks over a large time period. A market graph is constructed with nodes and edges; nodes represent the quantity of interest, or specific data points, such as stock prices at an instance of time. Edges represent a relationship between one node and another. Creating edges can be accomplished through many different approaches, including correlation coefficients, power law, and minimum spanning tree. Pearson's correlation coefficient can be used to relate a set of two data points and can be further filtered through a minimum threshold value. The power law graph is another unique way to relate data points to one another. The power law graph creates edges among nodes by considering a probability and the binomial distribution. The power law graph has powerful implications on network analysis because it concludes that the degree distribution, the number of connections a node has to other nodes, is represented as an exponential relationship. A minimum spanning tree is a hierarchical method used to analyze market graphs. A minimum spanning tree clusters data by partitioning data appropriately. Overall, many methods are defined to establish a market graph depending on the purpose of the analysis and the parameter of interest.

Introduction

The stock market is a financial market with fluctuations that provide immense data that can be used in a predictive model of market behavior. Further, having accurate ways to group and partition data is beneficial to establish relationships between different financial entities and to build trust in the analyses. Current techniques used to analyze large amounts of financial data primarily represent a large data set as a market graph. A market graph, visually, is a network consisting of nodes and edges. Connecting nodes can be accomplished with different algorithms.

One example that illustrates the powerful implications a market graph can have is in a study done by Chun-Xiao Nie. The study focused on the stock market in China versus the United States, while in a financial recession. By applying a correlation dimension and using five minute stock price data, Nie concluded that a correlation dimension could be used for analyzing market

volatility [4]. Thus, a network approach was able to explain the stock market's volatility during a financial recession.

Similarly, network approaches can be used on financial data, like stock prices, from a large data set to be able to further identify relations from one company to another. Examining methodologies that look at dependency, clustering, and density of a market graph will be beneficial to understanding conclusions drawn from these methods. Thus, the purpose of this paper is to review network approaches used to analyze the dynamics of financial markets.

Concept of Market Graph

Because financial data can be immense, hard to analyze, and hard to make connections or draw conclusions from, a network approach through the creation of a market graph is the ideal way to tackle this problem. Financial data can mean a variety of different things. For example, one can look at transactions, credit relations, stock prices, relations between different companies, etc. [1] One of the most basic ways to simplify data in order to draw conclusions is by creating a network graph. Network graphs are comprised of a set of nodes and a set of vertices/edges. In network graphs, nodes represent the quantity of interest; for example, a company's current stock pricing. The goal of a network graph is to be able to link the nodes with weighted edges. Each edge illustrates a relationship between one node and another. An edge, for example, can represent how relatively similar a company is to another company based on stock pricing or other characteristics [1].

Pearson's Correlation Coefficient

An issue that arises in graph theory is how to determine which nodes will be linked with an edge. A widely accepted tool to link nodes in a network graph is by using the Pearson's correlation coefficient [1]. Pearson's correlation coefficient is a linear correlation that measures similarity and is the best method of representing similarity-based networks.

To calculate Pearson's correlation coefficient, C_{ij}, one can use the formula given by:

$$C_{ij} = \frac{\{r_i r_j\} - \{r_i\}\{r_j\}}{\sqrt{\{r_i^2\} - \{r_i\}^2} * \sqrt{\{r_j^2\} - \{r_j\}^2}}$$

In this equation, r_i and r_j represent the characteristic being analyzed, for example stock prices. The brackets represent taking the expected value or an average of the specific data characteristic.

As an example, $\{r_i\} = \frac{1}{N}\sum_{k=1}^{N} r_k$, where N is the total number of nodes in the graph. Also, for reference, the denominator of the Pearson's correlation coefficient is the square root of the variance of node i's quantitative characteristic multiplied by the variance of node j's quantitative characteristic.

This formula produces, C_{ij}, which will be a value between [-1,1]. The calculated correlation coefficient will represent how correlated node i and node j are and can be repeated for any set of nodes. Once C_{ij} is calculated, a threshold value θ determines whether an edge is established between two nodes. If, $C_{ij} < \theta$, no edge is established between two nodes, while if $C_{ij} \geq \theta$, an edge is established between node's i and j.

Effects of a Threshold

As the threshold, θ, changes, the network graph changes as well. Determining an appropriate threshold is important to be able to accurately draw conclusions from a network graph. As discussed by Boginskia, Butenkob, and Pardalos, the smaller the threshold value, the less structure a network graph will have [2]. With a lower threshold, more edges are created and therefore more links between nodes are established. Also, as the threshold increases, the amount of edge connections a single node has (degree distribution) will follow the power law [2]. This concept is illustrated by higher threshold values being plotted as an almost straight line for a logarithmic scale [2].

A proposed way to find an optimal threshold value is presented by Battiston, Glattfelder, Garlaschelli, Caldarelli [1]. As stated, if each of the N total nodes are represented as an "independent Gaussian time series of length T…it is known that for large T the distribution of the sample correlation coefficient of two uncorrelated Gaussian variable can be approximated by a Gaussian distribution with zero mean and standard deviation equal $\frac{1}{\sqrt{T}}$" [1]. This is beneficial because depending on how conservative the market graph should be, there is a measurable way to determine an appropriate threshold. Typically, for example, three standard deviations is commonly accepted as a good confidence interval since in a Gaussian model 99.7% of the data will be attributed to being within three standard deviations. Thus, an edge linking two nodes is only placed if the correlation coefficient is larger than $\frac{3}{\sqrt{T}}$ [1]. Determining an appropriate threshold value is a significant factor in determining the structure of a market graph.

Power Law Graph Association

The power law graph association with a market graph further explains the relationship between establishing edges between nodes. A common result in constructing market graphs is the display of a heavy tail in the degree distribution of the market graph [3]. The degree distribution is defined as the amount of edge connections a single node has. This characteristic helps define what is known as a power law graph.

Erds and Renyi first introduced concepts of degree distribution [3]. They proposed creating edges between different nodes using an independent probability p. Therefore, by denoting a graph G, with p edges, and n nodes as G(n,p), Erds and Renyi concluded the degree distribution for any node to follow a binomial distribution would resemble: $P(d_n = k) = (\begin{matrix} n-1 \\ k \end{matrix})p^k(1-p)^{n-1-k}$ [3]. For a binominal distribution, as n approaches infinity, np becomes a constant, and takes on a Poisson distribution: $P(d_n = k) = \frac{(np)^k e^{-np}}{k!}$ [3]. The Poisson distribution illustrates why this type of distribution is called a power law distribution because of the exponential term in the numerator.

Specifically, the power law distribution is mostly found in the right end tail of a market graph's degree distribution. A random variable, X, is said to follow the power law if its probability density function is of the form: $f(x)_x = \frac{A}{X^B}$, where A is a normalization constant and B is the power law exponent [3]. Thus, a power law graph is formed if the number of nodes n with a degree of x, follows: $n = \frac{e^A}{x^B}$ [3]. The power law model is a powerful formulation because it directly concludes a statement relating any node and its degree distribution by representing this relationship exponentially [2].

Minimum Spanning Tree

A beneficial tool when analyzing market graphs is to look for clusters and sub clusters which can lead to a hierarchical structure in the data [1]. Clustering of data consists of partitioning data into groups, based on having a certain degree of similarity within a group and a certain degree of difference between one cluster of data and another. Another important rule when partitioning data is if there is not a link between two nodes/vertices, then those two nodes may not be placed

in the same cluster in order for each cluster to be statistically relevant and to be of high quality [3].

One algorithm that accomplishes displaying a data's hierarchical structure is the minimum spanning Tree procedure. The minimum spanning Tree, a hierarchical analysis method, entails building a graph by first establishing a distance matrix where each entry, d_{ij}, is equal to $\sqrt{2(1-C_{ij})}$ [4]. The goal of MST is to minimize the distance, d_{ij} , that links node i to node j, or alternatively maximize C_{ij}, the Pearson's correlation coefficient associated with linking node i and j. Each link whether it be d_{ij} or C_{ij} represents a "length" connecting two nodes together. The minimum spanning tree visually can show how similar nodes are because similar nodes will be placed close together and therefore have a small distance, d_{ij}, length between them. One procedure used to retrieve the minimum spanning tree is suggested by Battiston, Glattfelder, Garlaschelli, and Caldarelli [1]:

Minimum Spanning Tree Procedure [1]:

1. Assign distances between vertices so that the shortest is the distance and the largest is the correlation between two vertices.
2. Rank these distances from the shortest to the longest.
3. Start from the shortest distance and "draw" the edge between the vertices.
4. Iterate until you find an edge that would form a loop. In this case, jump to the next distance (if necessary repeat this operation).
5. Stop when all the vertices have been considered.

Once completed, a minimum spanning tree will consist of N-1 links [1]. Other algorithms like the minimum spanning tree include: average linkage minimum spanning tree which can remove unwanted chaining associated with MST, planar maximally filtered graph, Directed Bubble Hierarchical Tree, Triangulated Maximally Filtered Graph, and many other approaches which can combat issues with MST and can focus in on other parameters [5].

An issue that arises under the minimum spanning tree procedure is displayed in Figure 1 [6]. Nodes 3 and 5 have a large weight of 6, but because the correlation is lower than between nodes 2 and 3, the link between nodes 3 and 5 is removed [7]. Intuitively, the higher weight between nodes 3 and 5 would be more beneficial to have kept in the new graph. The deleted link between nodes 3 and 5 creates a different conclusion from the data than if the link still existed. Therefore, the MST procedure can affect the true topology of a graph [7].

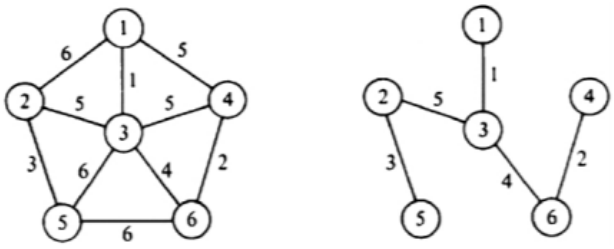

Figure 1. Left image displays original graph while the right image displays the graph with the minimum spanning tree procedure applied [6]

Cliques and Independent Sets

A clique within a network graph is a set of interconnected nodes and vertices while an independent set is a set of nodes/vertices that do not have connections [8]. To find the largest clique, a maximum clique problem can be set up and similarly to find the largest independent set a maximum independent set problem can be established. The cliques and independent sets are established from a market graph that would connect vertices based on similarity, like the Pearson's correlation coefficient. Thus, a clique will be useful because it will represent a dense cluster of similar objects [2]. On the other hand, independent sets are the opposite as they will represent groups of objects that differ from other objects within the group.

In practice, a clique is a powerful tool to draw conclusions from data. For example, a clique with a positive threshold value will represent a set of instruments whose price fluctuations have a similar behavior [2]. All data values in a clique will be considered similar to one another, where if one changes another data value will in turn also change. In independent sets in a market graph, if the threshold value is negative, the conclusions drawn are that the financial instruments grouped together will be negatively correlated to one another and therefore are not related, which is considered a completely diversified portfolio [2].

Maximum clique and independent set problems are beneficial because they provide the maximum possible size of groups containing either similar or different objects.
Bomze, Budinich, Pardalos, and Pelillo formulated the maximum clique problem as an integer programming problem [8]. To use the procedure, first there must be an established large clique in the market graph that will represent the lower bound of the size of the maximum clique. To establish an initial large clique a greedy heuristic algorithm can be used which creates a clique through "recursively adding a vertex from the neighborhood of the clique adjacent to the most vertices in the neighborhood of the clique" [2]. To have the correct scope of data, all vertices

with a degree less than the size of the clique found in the greedy heuristic algorithm are recursively removed from the overall market graph. This process removes data values that would not fit correctly in a maximum clique problem.

The integer formulation of the maximum clique problem is:$\max \sum x_i \;\; s.t. \; x_i + x_j \leq 1$, $(i,j) \notin E, x_i \in \{0,1\}$, where E represents the edges kept after limiting the data scope [8]. A result of this algorithm is that as a positive threshold value increases, clique sizes decrease because there is a harder constraint on what classifies a node similar to another node [2]. The maximum independent set problem, on the other hand, is harder to solve. The results, similar to the maximum clique problem, display that as a negative threshold value places a harder constraint on similarity, the independent clique size decreases [2]. Another important note is that the maximum independent set problem results in smaller set sizes than the maximum clique problem because of the added difficulty that comes with solving an independent set problem. This implies that finding a completely diversified portfolio is not an easy task in the stock market because the independent clique, the number of negatively correlated financial instruments, is small [2]. Overall, the maximum clique and independent set problem can be a beneficial tool for investors because they can see which assets can build a diversified portfolio to reduce investor risk or which assets will move together, which can also be beneficial for an investor.

Conclusion

The three most common network approaches used to analyze large sets of financial data are Pearson's correlation coefficient, the power law graph association, minimum spanning tree, and cliques and independent sets. Pearson's correlation coefficient is a common method used because the market graph can be filtered by different threshold values, thus affecting the structure of the graph. Also Pearson's correlation coefficient is used for similarity based networks, for example when comparing data between two financial instruments.

The power law graph is an important method because it implicates an exponential relationship for a network's degree distribution. The minimum spanning tree draws a network by using distances and correlations between nodes to cluster data into different partitions. Partitioning data has powerful implications because each cluster of data must have a certain degree of similarity and differing clusters as well must have a certain degree of difference. Thus, the

minimum spanning tree can provide a market graph that can show how data may be clustered. Cliques also collect a group of financial instruments that are positively correlated to each other, which can help investors understand how stock prices of one company may change the price of another company. Independent sets, while harder to calculate than cliques, are useful for collecting negatively correlated financial instruments. Independent sets, while typically smaller in size, are useful in making a diversified portfolio.

References

[1] Battiston, S., Glattfelder, J., Garlaschelli, D. and Caldarelli, G. (n.d.). The Structure of Financial Networks. [Online] Available at: https://www.sg.ethz.ch/media/publication_files/chp3A10.10072F978-1-84996-396-1_7.pdf.

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

[3] K. Sorensen and P. Pardalos, "Clustering in Financial Markets." In: Kalyagin V., Koldanov P., Pardalos P. (eds) Models, Algorithms and Technologies for Network Analysis. NET 2014. Springer Proceedings in Mathematics & Statistics, vol 156. Springer, Cham. 2016.

[4] C.-X. Nie, “Correlation dimension of financial market,” Physica A: Statistical Mechanics and its Applications”, vol. 473, pp. 632–639, May 2017.

[5] G. Marti, F. Nielsen, M. Bińkowski, and P. Donnat, “A review of two decades of correlations, hierarchies, networks and clustering in financial markets.” eprint arXiv:1703.00485, 03/2017.

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

[7] N. Huang, L. Leo Keselman, and V. Vincent Sitzmann, “Beyond correlation networks.” http://snap.stanford.edu/class/cs224w-2016/projects/cs224w-73-final.pdf

[8] Bomze 1999 I.M. Bomze, M. Budinich, P.M. Pardalos, M. Pelillo. The maximum clique problem D.-Z. Du, P.M. Pardalos (Eds.), Handbook of Combinatorial Optimization, Kluwer Academic Publishers, Dordrecht (1999), pp. 1-74