Optimizing Stock Return Predictions Using Network Centrality Measures Based on Maximized Lagged Correlations

Joy Batty
Summer, 2024
joylbatty@lewisu.edu

Abstract

This project aims to enhance stock return predictions by leveraging network centrality measures based on maximized lagged correlations between stocks. Historical stock data was retrieved, processed into weekly returns, and used to compute correlation matrices with optimal lag periods. A network representing stock relationships was constructed, with centrality measures calculated to quantify each stock's importance.

Long Short-Term Memory (LSTM) networks, a type of recurrent neural network (RNN), were employed in this study due to its ability to handle long-term dependencies making it particularly well-suited for time series prediction. LSTMs can capture temporal dynamics in data, which is crucial for accurate forecasting of stock returns.

Two models were developed for comparison: one utilizing centrality measures as exogenous variables and another employing autoregression based solely on past returns. The LSTM model demonstrated its ability to effectively incorporate these complex features, achieving a SMAPE of 11.35% and an RMSE of 28.26, compared to 11.66% and 28.99 without centrality measures. This improvement, though modest, highlights the potential of centrality measures to enhance predictive accuracy by capturing the intricate relationships within the stock market network. In contrast, the Linear Regression model performed poorly with centrality measures (SMAPE: 41.70%, RMSE: 77.08), but improved significantly without them (SMAPE: 11.55%, RMSE: 30.77). These findings underscore the importance of using appropriate models to leverage the benefits of centrality measures as exogenous variables, demonstrating that while LSTM models can utilize advanced

network analytics for better predictions, simpler models like Linear Regression may perform better with straightforward, direct features.

This analysis underscores the crucial role of model selection and feature engineering in financial predictive analytics, highlighting the potential of LSTM models to effectively leverage complex features like centrality measures. These findings open the door to further enhancements and optimizations, paving the way for even greater improvements in accuracy. The project showcases the promise of advanced network analytics in financial forecasting, providing a solid foundation for future research to build upon and refine these innovative techniques.

I. INTRODUCTION

Financial markets are denoted by their complexity and interdependence, where the performance of one asset often influences others [1]. One of the key methods to analyze these interdependencies is through the study of correlations between stock prices [2]. Understanding these relationships can be critical for making informed investment decisions.

Correlations between stock prices can be used to construct networks. Network analysis has become a more prevalent tool for understanding stock market dynamics [3]. Each node represents a stock and edges between nodes represent the strength of the correlation between the respective stock prices [4]. For instance, a high positive correlation between two stocks would be represented by a strong edge, indicating that these stocks tend to move in the same direction . Conversely, a negative correlation would suggest that as one stock's price increases, the other's decreases, which can be depicted by an edge of a different nature [5]. By representing stocks as nodes and their relationships (such as correlations or transactions) as edges, network analysis provides a comprehensive view of the interdependencies within the market [6]. This method allows researchers and investors to visualize and quantify the complex web of interactions that drive market behavior [7].

An important aspect of analyzing correlations is considering the lag between the movements of different stocks. Lagged correlations account for the time delay between the price movements of different assets, which can provide deeper insights into the lead-lag relationships within the market [8]. For example, if the price of one stock tends to rise a few days before another, this lead-lag relationship can be captured by computing the correlation with a time lag [9].

By incorporating lag into the correlation analysis, we can construct more dynamic and informative networks. These lagged correlation-based networks reveal not just simultaneous relationships but also how past movements of one stock can influence future movements of another. This approach can help identify leading stocks that can be early indicators of market trends and lagging stocks that follow the market [10].

Once the network is constructed, it can be analyzed using various network metrics. For example, centrality measures such as betweenness, closeness, and eigenvector centrality can identify key stocks that play pivotal roles within the market [11]. Network metrics can also be employed for prediction purposes. For instance, stocks with high centrality measures may be used as predictors for future market movements due to their influential roles within the network. Researchers have found that incorporating network features such as centrality measures into predictive models can enhance the accuracy of forecasts [12]. This is because these features capture the essential structural properties and dynamics of the market, providing a more comprehensive set of explanatory variables for predictive modeling [13].

II. Discussion of Related Work

The below information reviews previous research which pertains to this project. Many researchers have explored various methods for constructing networks to analyze stock correlations, and for determining exogenous variables in order to build machine learning models for stock return prediction.

A. Network Construction Using Correlation Weights

In the study "A Network Approach to Portfolio Selection," the authors utilize different correlation weights to construct networks representing the relationships between assets in the financial market . They then analyze various centrality measures within these networks to understand the importance and influence of individual stocks within the network structure. The authors use correlation weights derived from stock returns to construct networks where securities are represented as nodes and the links between them reflect the correlations of their returns. By varying the correlation weights (e.g., considering correlations at different lags), they create different network structures that capture the interconnections between assets. After constructing the networks based on correlation weights, the study examines centrality measures within these networks. Centrality measures, such as degree centrality, betweenness centrality, or eigenvector centrality, provide insights into the importance or influence of individual securities within the network. [14]

B. Network Construction Using Pearson Correlation and VAR Model

Both the Pearson correlation and VAR model are used to construct volatility networks. These networks reflect the relationships between the stock indices based on their volatilities. Network indicators such as centrality measures are calculated to capture the importance and influence of each node (stock index) within the network. These indicators are then used as inputs for machine learning models to predict market directions and to design investment strategies [15].

C. Network Construction Using Mutual Information and Big Data

The authors constructed their stock correlation networks using mutual information and financial big data. They collected stock price data for a specific period and calculated mutual information and correlation coefficients for each pair of stocks. Using the Minimum Spanning Tree (MST) method, they constructed an undirected weighted network based on these values. The network was visualized with different colors for stocks from various sectors, highlighting both internal and inter-sector connections. Analysis of the network's structure, clustering patterns, and topological properties provided insights into the relationships between stocks and sectors, aiding in understanding market dynamics and potential future price predictions [16].

D. Network Construction Using Correlation Weights

The study aimed to construct a network capturing strong lead-lag relationships within a multivariate time series system. By identifying and representing these relationships as directed edges, the researchers analyzed the overall properties of the lead-lag connections using network analysis tools. They focused on detecting clusters of variables with significant lead-lag behavior. The network was constructed with nodes representing time series variables and directed edges indicating lead-lag relationships, weighted by the relationship's magnitude. They then used directed network clustering to partition the system and quantify the leadingness of each cluster. [17]

E. Time-Lag Selection

In the realm of time-series forecasting, selecting the optimal lag period is crucial for enhancing predictive accuracy. The paper "Time-Lag Selection for Time-Series Forecasting Using Neural Network and Heuristic Algorithm" by Surakhi et al. provides a compelling example of this. The researchers utilized neural networks and genetic algorithms to identify the lag period that maximized the accuracy of their forecasts. This methodology can be particularly relevant in financial contexts, where understanding the temporal dependencies between past and future prices can significantly improve prediction models [18].

F. Constructing Networks Using Mutual Information for Machine Learning Prediction

The researchers gathered time-series stock price records of the S&P 500 underlying companies and split these records into one-hour intervals. They calculated mutual information for stock pairs based on price movements and used this information to construct networks with mutual information as link weights. They computed the node strength distribution for each network and built metrics using strength distribution, network centrality, and modularity. The researchers then created linearly combined predictors using the two best-performing metrics and built ARIMA models to predict the actual S&P 500 index. Finally, they evaluated whether network measurements improved the accuracy of the ARIMA models. This process allowed them to analyze the interdependencies and dynamics of the S&P 500 companies using network science techniques [19].

G. Using Arima to Predict Stock Movement

The paper "Forecasting Multinomial Stock Returns Using Machine Learning Methods" by L. Nevasalmi built a data library of 43 corporate bond-level characteristics based on existing literature on the cross-section of corporate bonds which they used as predictors. These predictors are designed to cover a broad set of corporate bond return predictors, including bond-level characteristics, risk proxies, bond-level illiquidity measures, past bond return characteristics, and distributional characteristics. Additionally, they combined these bond-level characteristics with 94 stock characteristics used in previous studies [20].

III. Project Description

This project aims to optimize stock return predictions by leveraging network centrality measures based on maximized lagged correlations between stocks. Initially, historical stock data was retrieved from Yahoo Finance and processed to calculate daily returns. These daily returns were then aggregated into weekly returns based on a Thursday-to-Wednesday close period over a span of five years. This aggregation was done to facilitate the computation of weekly correlation matrices.

To construct a network representing the relationships among stocks, the maximum correlation for each stock pair was determined by identifying the optimal lag, with a lag period of up to 4 weeks being used to get the maximum correlation. Additionally, p-values from the maximum correlation matrices were used to ensure the statistical significance of the correlations. The network graph was built where nodes represented stocks and edges were weighted by the maximum correlations. Centrality measures, such as PageRank, betweenness centrality, and closeness centrality, were then computed for each stock to quantify its importance within the network. These centrality measures were visualized for each company over time to identify trends and patterns. Additionally, the correlation between these centrality measures and the stock returns was analyzed to understand their predictive power.

To evaluate the effectiveness of including network centrality measures as predictive features, two models were developed: one utilizing centrality features as exogenous variables and another using plain autoregression. By comparing the performance of these models using metrics such as Symmetric Mean Absolute Percentage Error (SMAPE), the project aimed to demonstrate the potential improvements in prediction accuracy brought by incorporating network analysis. This comprehensive approach not only provides insights into the interconnected nature of stock returns but also highlights the value of advanced network analytics in financial forecasting.

Researchers have used many techniques to construct their networks and for predictor selection. The novelty of this project lies in using centrality measures derived from networks where the edges are weighted based on the lagged correlations that maximize the correlation between stock pairs as exogenous variables for prediction. By identifying the optimal lag for each pairwise correlation and using these centrality measures in predictive models, the project aims to improve the accuracy and robustness of stock return predictions.

IV. Methodology

The project can be broken down into three parts. Part 1 focuses on calculating various lagged correlations between stock pairs and determining the lag that maximizes correlation. Part 2 involves constructing a network using these maximized correlations as edge weights. Part 3 entails using the centrality measures derived from this network as variables in machine learning models to predict stock returns. This structured approach aims to leverage optimized correlations and network centrality to enhance predictive accuracy in stock market forecasting.

Part 1: Data Collection, Processing, and Correlation Calculation

1. Data Collection:

- Objective: Gather daily stock price data over a 5 year period and process it.
- Sources: Yahoo Finance.

Method:

- Calculate the average volume for the 5 years for each company, and filter for the top 50 companies.
- Group daily returns into weeks.
  - Thursday close to Wednesday close.

2. Calculation of Lagged Correlations:

- Objective: Identify the lag that maximizes correlation for each pair of stocks.
- Method:
  - Calculate weekly lagged correlations with a lag period of 4 weeks.
  - Construct an adjacency matrix where each element represents the maximized correlation between a pair of stocks.

a. *Part 2: Network Construction and Centrality Measures*

b. 3. Network Construction:

- Objective: Identify Correlation Data
- Method:
    - Use the maximum correlations calculated in Part 1 to create a combined directed and undirected weighted network.
        - Lag = 0 is undirected
        - Lag > 0 is directed
    - Nodes represent stocks, and edges represent the correlation between stock pairs, weighted by the correlation value.

c. 4. Calculation of Centrality Measures:

- Objective: Compute centrality measures for each node.
- Measures: PageRank, betweenness, closeness, eigenvector centrality.
- Tools: Python libraries like NetworkX.
- Calculate correlation values between the centrality measures and the returns.
- Calculate the p-values to determine the statistical significance of these correlations.
- Generate time series plots for each centrality measure, showing their variation over time for each stock.

Part 3: Predictive Modeling

5. Predictive Modeling:

- Objective: Use centrality measures to predict stock returns.
- Models: Train a Last Short Term Model (LSTM)
    - Train a LSTM model using centrality measures.
    - Train a baseline autoregressive model using past returns.

- ○ Evaluation: Use sMAPE and RMSE.
- ○ Compare the performance of both models.

V. Implementation Description

B. Yahoo Finance Data Fetching

Historical stock data was retrieved for a list of common stock symbols from the NYSE utilizing the yfinance library. It should be noted that it would be best to filter out Exchange Traded Funds (ETF's) at this point, however for this project, it was done at a later stage. The collection range is from January 1, 2019, to January 1, 2024.

C. Determine Top 50 Companies

The top 50 companies were filtered based on average trading volume for each company using the groupby and mean functions from pandas. This average volume is used to sort the companies in descending order, effectively prioritizing the most actively traded stocks. To facilitate weekly return analysis, the daily returns of these 50 companies were grouped into weekly groups, from Thursday to Wednesday.

D. Weekly Correlation and Lag Calculation

Weekly correlations were calculated between stock returns with different lag periods and stored the results in correlation matrices and optimal lag matrices. The script utilizes pandas for data manipulation, numpy for numerical operations, and scipy.stats for calculating Pearson correlation coefficients. For each pair of stocks, the function calculates the correlation between their returns for different lag periods, ranging from -4 to 4. The maximum correlation value and the corresponding lag period was tracked. This structured approach enables detailed analysis of the lead-lag relationships between stock returns, facilitating financial predictive analysis. P-values were also calculated for each correlation matrix which will be used for network construction.

E. Network Construction

A graph network was constructed with combined directed and undirected edges based on correlation thresholds (0.8) and p-value criteria (.05). Directed edges correspond to lags greater than 0, and undirected edges correspond to lag 0. Edges were also weighted using the correlation values. Centrality measures, including degree centrality, betweenness centrality, closeness centrality, PageRank, and eigenvector centrality, are computed for each network. Next, the correlation between the centrality measures and the weekly returns were computed. P-values were calculated from these correlation matrices. To provide visual insights, time series plots for each centrality measure, showing their variation over time for each stock was generated. These visualizations help to observe trends and potential relationships between centrality measures and stock performance.

F. LSTM Construction

Long Short-Term Memory (LSTMs) are a type of recurrent neural network (RNN) which can be used with sequence prediction problems. To create the LSTM I followed, Time Series Prediction with LSTM Recurrent Neural Networks in Python with Keras by Jason Brownlee.

The final part of the project focuses on analyzing the effectiveness of using significant centrality measures in predicting stock returns using an LSTM model, compared to a baseline autoregressive model that relies solely on past returns. These centrality measures are normalized and used as features for training an LSTM model. The stock returns were also normalized with a look-back period of 4 a forward window of 1 to capture time-series dependencies. Each stock will have its separate model.

The LSTM model is trained and its performance is evaluated on both training and test sets using RMSE and sMAPE metrics. To establish a baseline, the script also trains a separate LSTM model using only past returns as inputs. The performance of this autoregressive model is similarly evaluated and compared to the centrality-based model. This comprehensive approach allows for a detailed assessment of whether incorporating centrality measures offers any predictive advantage over traditional autoregression methods.

VI. Results and Analysis

To better understand the stock market fluctuations over the five-year period, Figure 1 below illustrates the network for the first week (1-10-2019), while Figure 2 depicts the network for the final week (12-28-23).

a. A. Network Visualizations

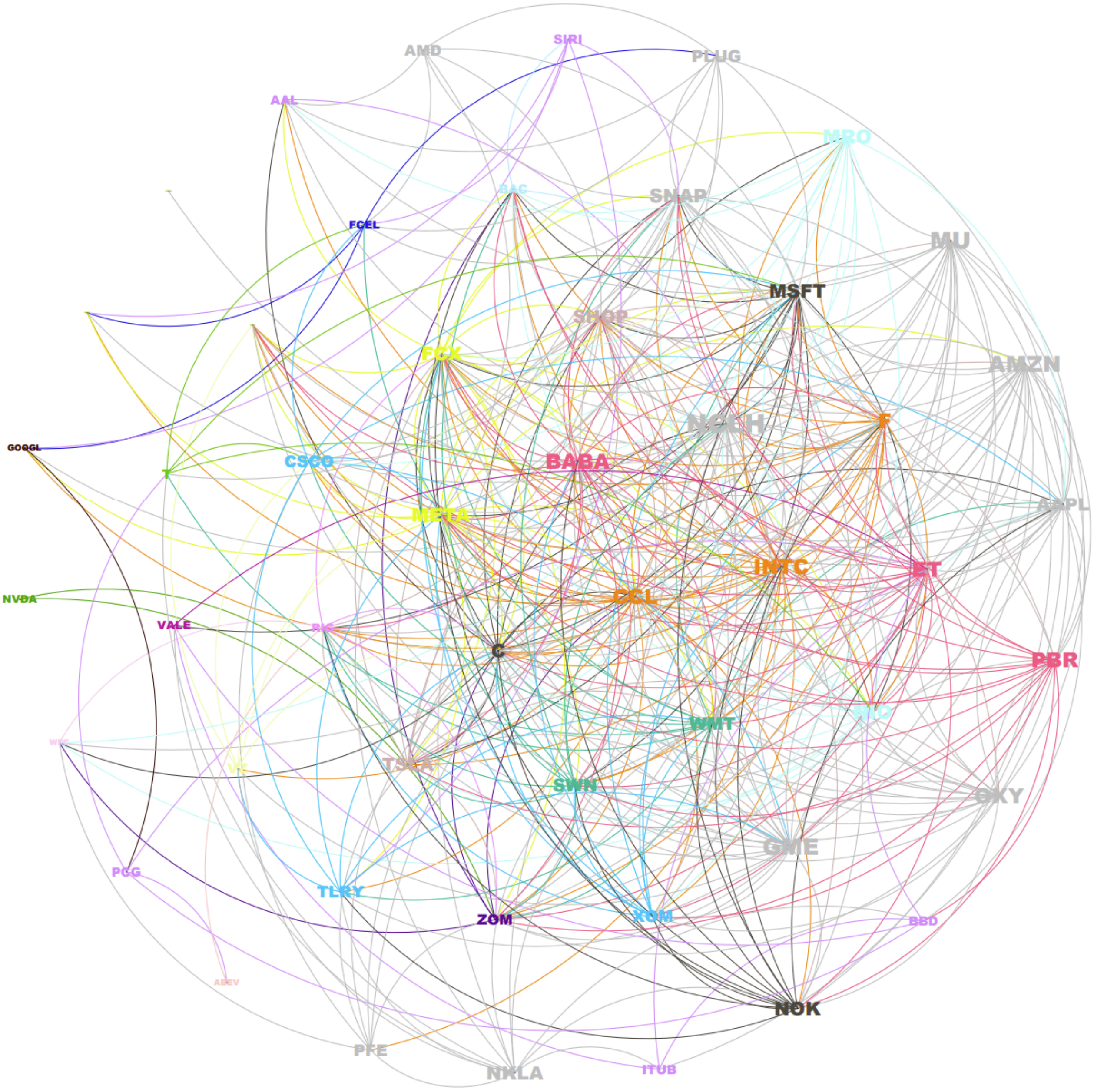


Fig. 1 Network for the week of 1-10-2019.

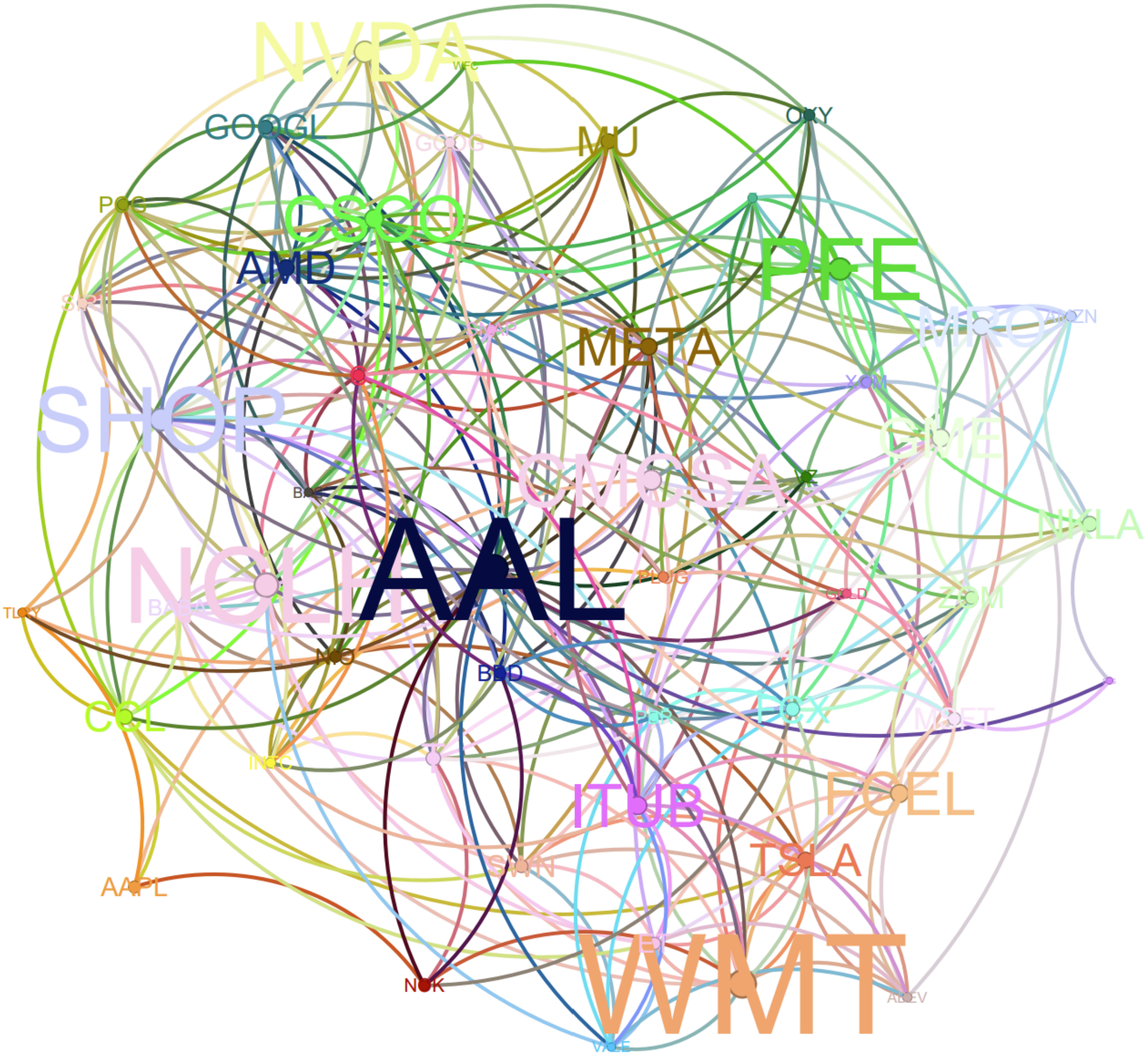


Fig. 2 Network for the week of 12-28-23.

Figures 1 and 2 above were constructed to show the network for the 50 companies for the first and last week of the dataset. The nodes were partitioned by Betweenness. Nodes with higher scores have larger fonts.

b. B. Centrality Measures Analysis

Table 1
Closeness Centrality 1-10-2019

Node	**Closeness**
NCLH	0.652777
MU	0.610389
GME	0.602564
AMZN	0.594936
OXY	0.573170

NCLH - Norwegian Cruise Line Holdings Ltd.

MU - Micron Technology, Inc.

GME - GameStop Corp.

AMZN - Amazon.com, Inc.

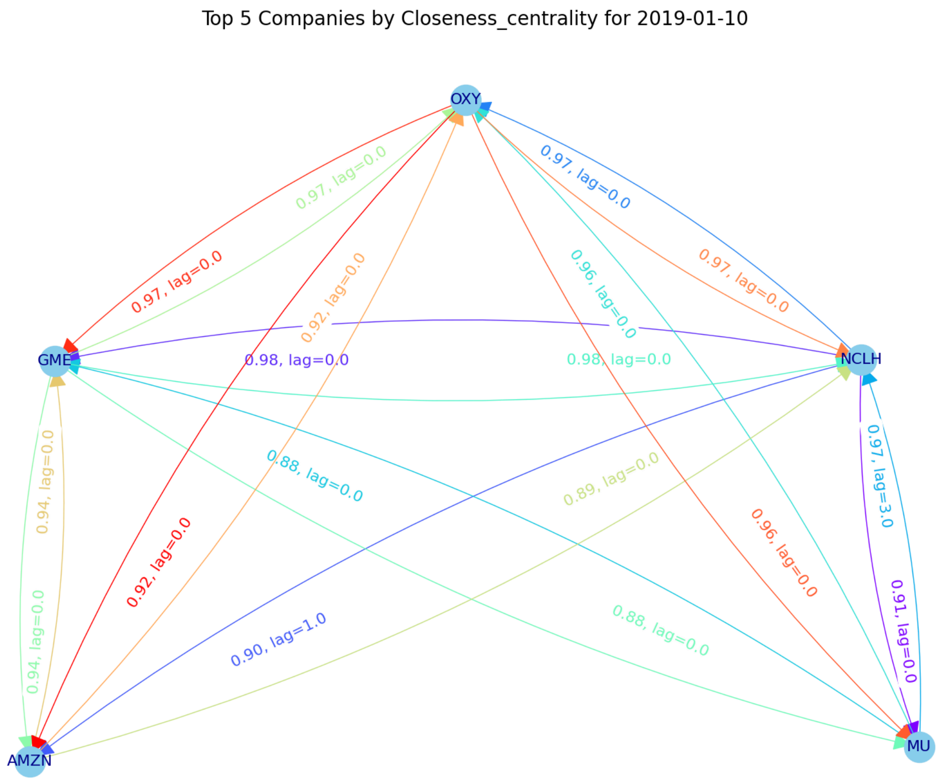


Fig. 3 Top 5 Companies by Closeness for 1-10-2019

Table 2
Betweenness Centrality 1-10-2019

Node	**Betweenness**
ET	245.265803
CCL	182.699830
SNAP	182.438229
C	180.0650731
AAPL	178.7350704

ET - Energy Transfer LP

CCL - Carnival Corporation & plc

SNAP - Snap Inc.

C - Citigroup Inc.

AAPL - Apple Inc.

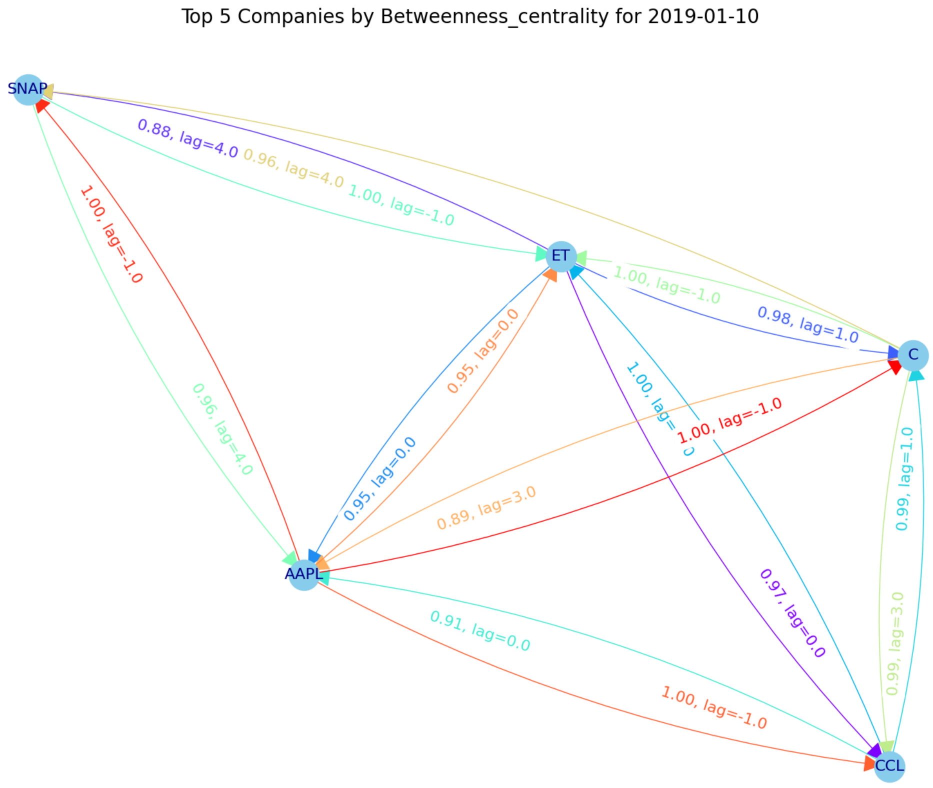


Fig. 4 Top 5 Companies by Betwenness for 1-10-2019

Table 3
PageRank Centrality 1-10-2019

Node	Pagerank
AAPL	0.02411971946479565

TSLA	0.023767154802140332
SHOP	0.02354619920426855
CSCO	0.023442788140138508
SWN	0.022476537868412198

AAPL - Apple Inc.

TSLA - Tesla, Inc.

SHOP - Shopify Inc.

CSCO - Cisco Systems, Inc.

SWN - Southwestern Energy Company

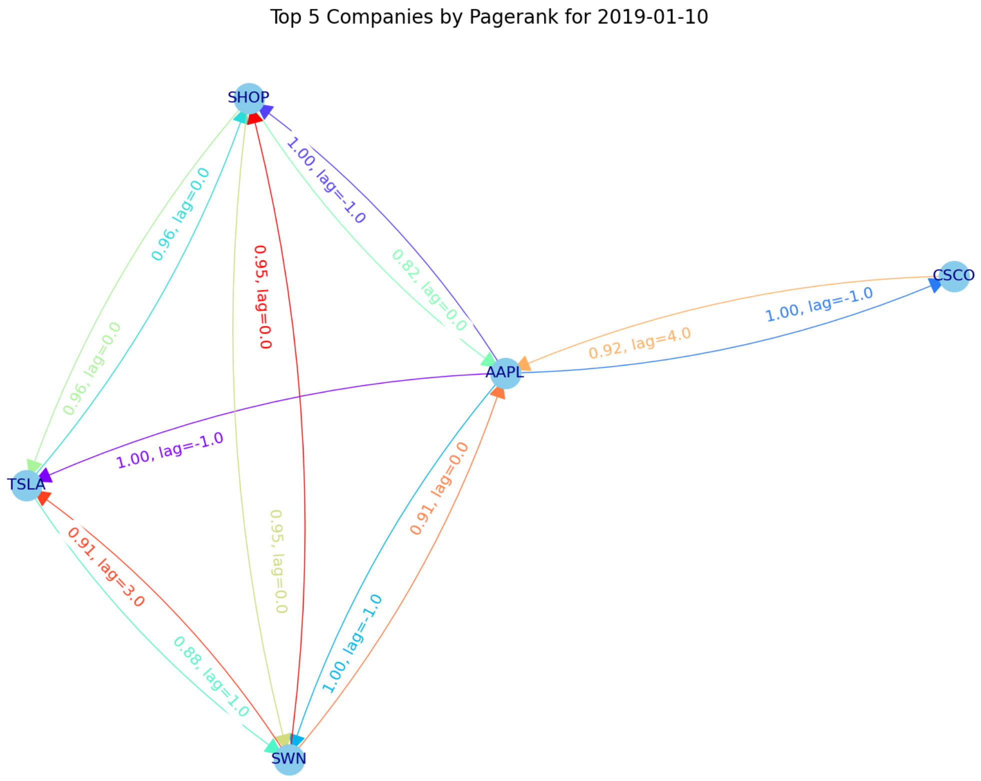

Top 5 Companies by Pagerank for 2019-01-10

Fig. 5 Top 5 Companies by PageRank for 1-10-2019

Table 4
Eigenvector Centrality 1-10-2019

Node	Eigenvector
SWN	0.5438554752527294
SHOP	0.5257438117751383
RIG	0.5165413769691756
NCLH	0.4724282577632497
TSLA	0.4599488910814239

SWN - Southwestern Energy Company

SHOP - Shopify Inc.

RIG - Transocean Ltd.

NCLH - Norwegian Cruise Line Holdings Ltd.

TSLA - Tesla, Inc.

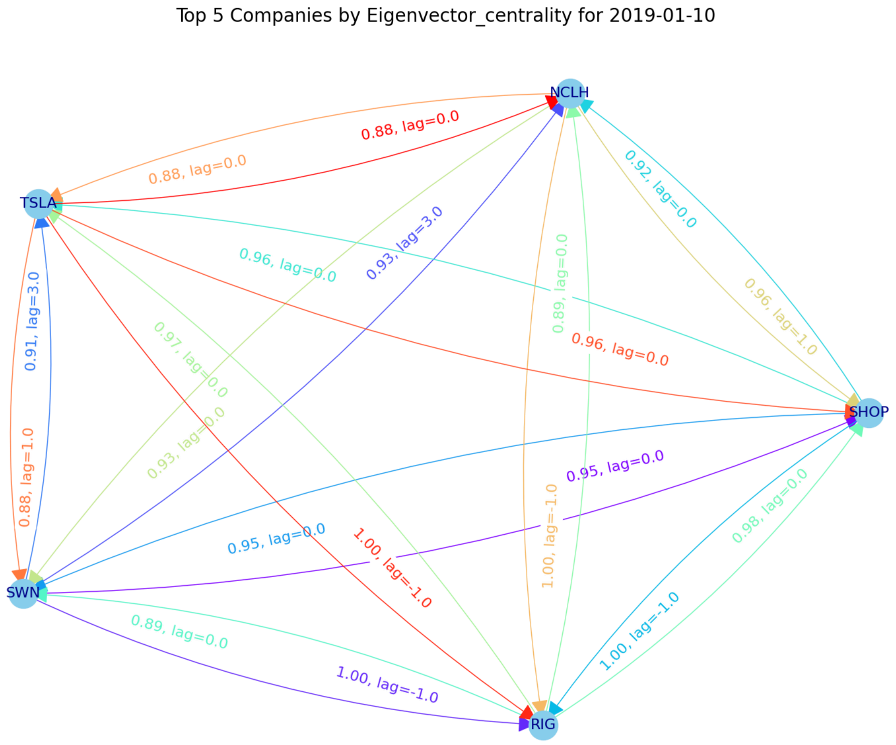


Fig. 6 Top 5 Companies by Eigenvector for 1-10-2019

Table 5
Closeness Centrality 12-28-23

Node	Closeness
NVDA	0.510638
MU	0.484848
BAC	0.475248
ITUB	0.475248
NCLH	0.470588

NVDA - NVIDIA Corporation

MU - Micron Technology, Inc.

BAC - Bank of America Corporation

ITUB - Itaú Unibanco Holding S.A.

NCLH - Norwegian Cruise Line Holdings Ltd.

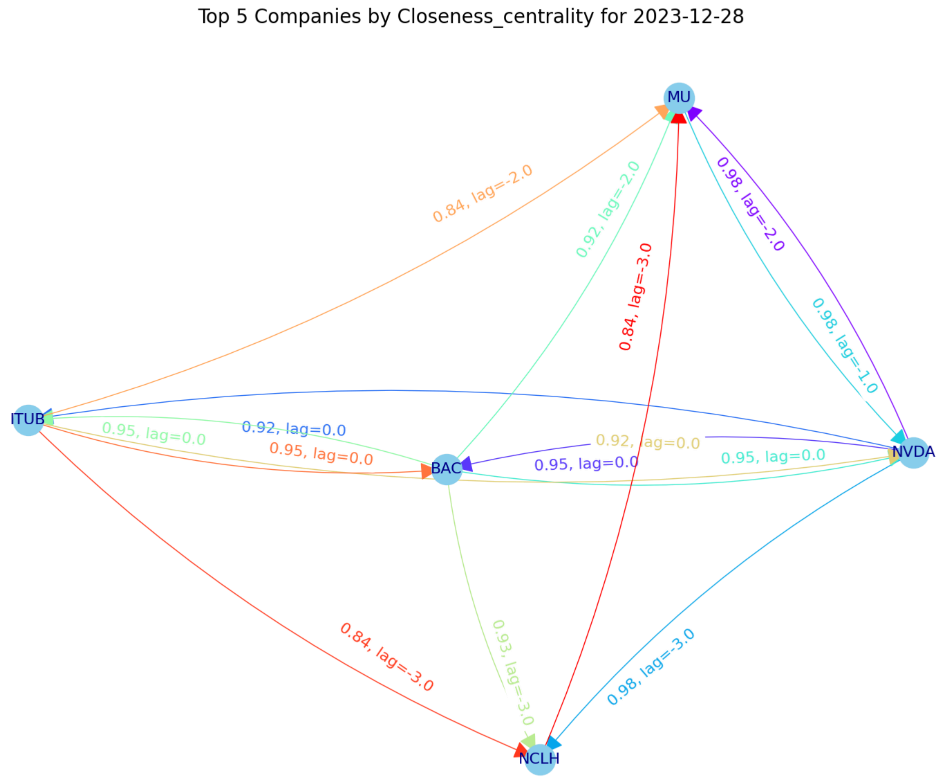


Fig. 7 Top 5 Companies by Closeness for 12-28-2023

Table 6
Betweenness Centrality 12-28-23

Node	**Betweenness**
AAL	216.162221
WMT	210.536513

NCLH	175.108807
SHOP	157.338233
PFE	152.654125

AAL - American Airlines Group Inc.

WMT - Walmart Inc.

NCLH - Norwegian Cruise Line Holdings Ltd.

SHOP - Shopify Inc.

PFE - Pfizer Inc.

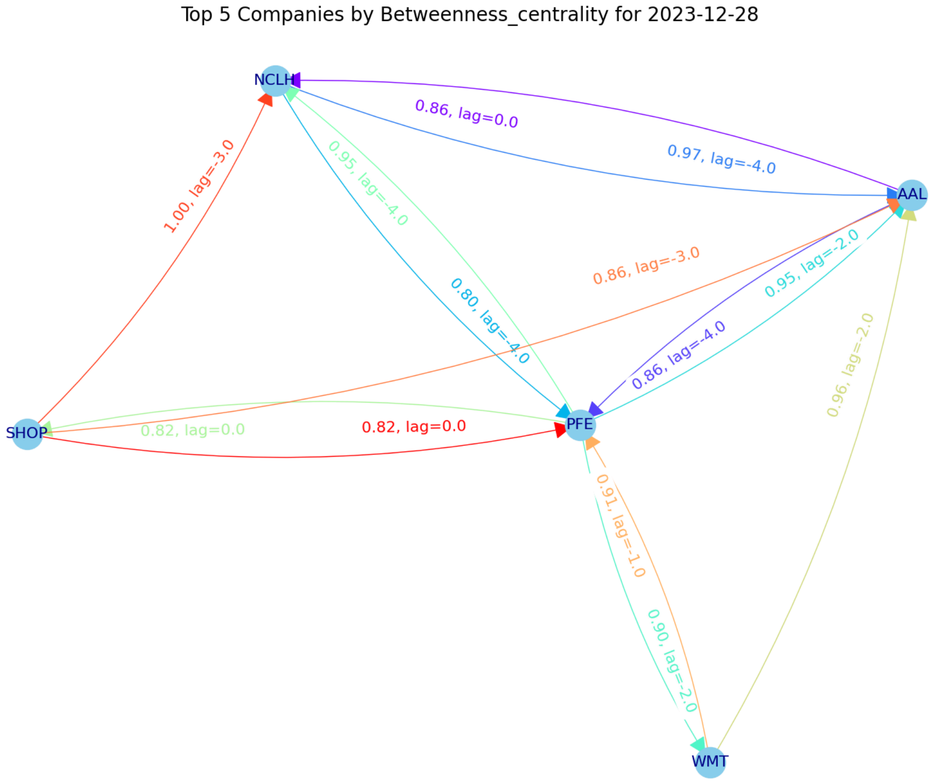


Fig. 8 Top 5 Companies by Betweenness for 12-28-2023

Table 7
PageRank Centrality 12-28-2023

Node	**Pagerank**
AAL	0.052097
NCLH	0.039466
WMT	0.038069
MSFT	0.037744
BABA	0.037329

AAL - American Airlines Group Inc.

NCLH - Norwegian Cruise Line Holdings Ltd.

WMT - Walmart Inc.

MSFT - Microsoft Corporation

BABA - Alibaba Group Holding Limited

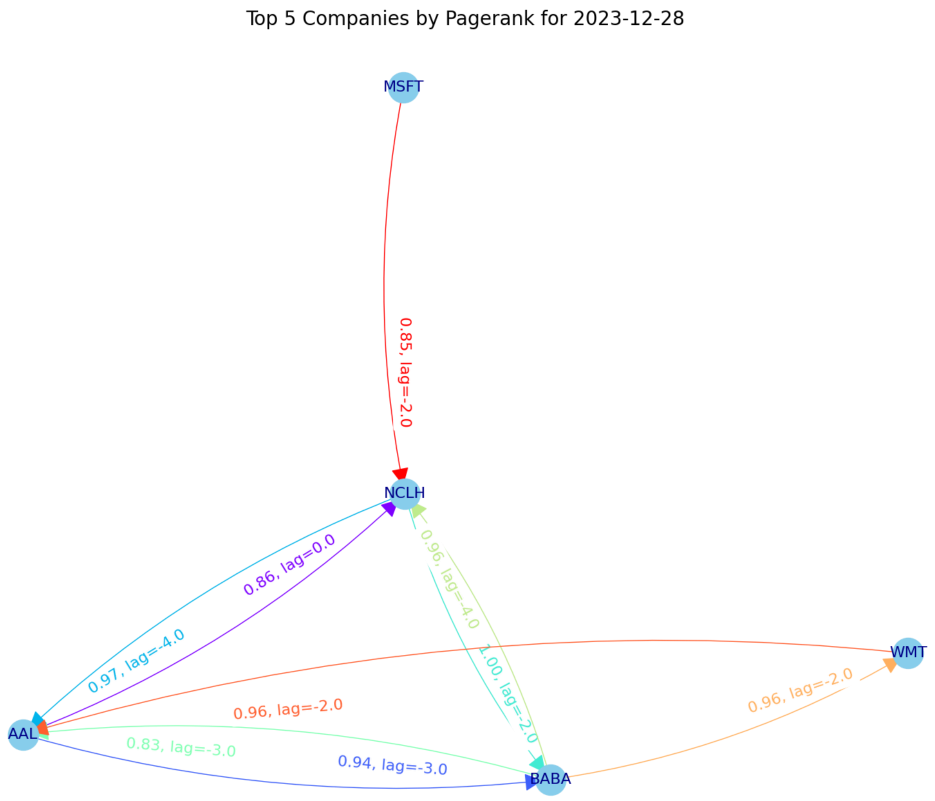


Fig. 9 Top 5 Companies by PageRank for 12-28-2023

Table 8
Eigenvector Centrality 12-28-2023

Node	**Eigenvector**
AAL	1.000000
C	0.938214
AMD	0.827448
PCG	0.794416
T	0.789825

AAL - American Airlines Group Inc.

C - Citigroup Inc.

AMD - Advanced Micro Devices, Inc.

PCG - PG&E Corporation

T - AT&T Inc.

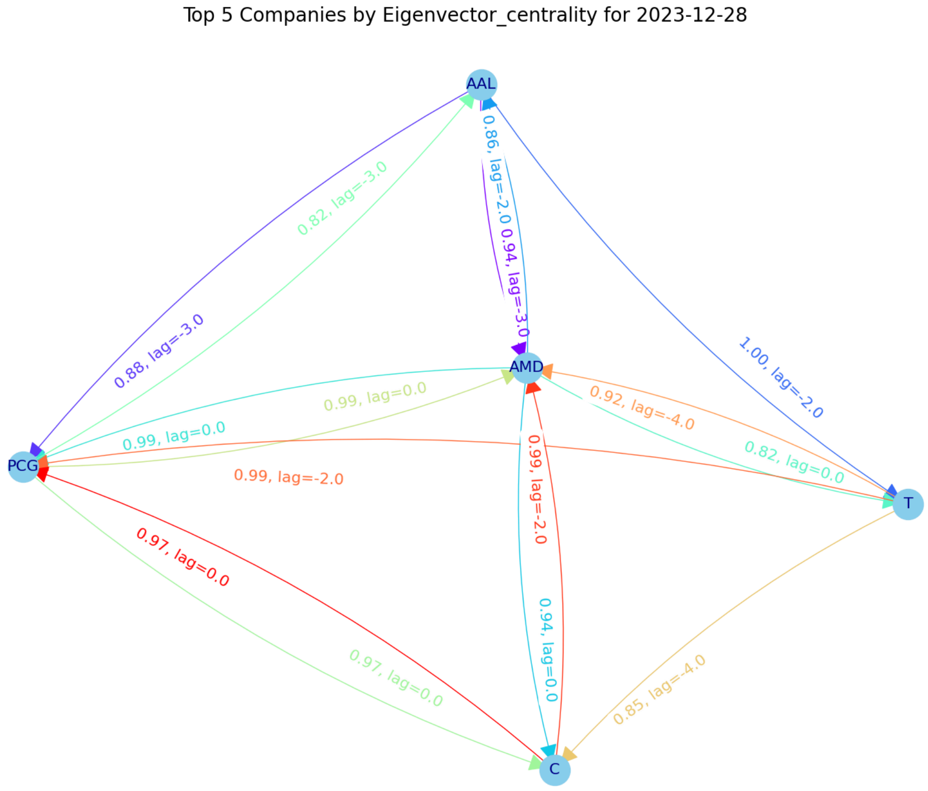


Fig. 10 Top 5 Companies by Eigenvector for 12-28-2023

In order to evaluate how the centrality measures have changed over time, tables 1-4 above aggregate the centrality measures for the first week of the dataset whereas tables 4-8 are the last week of the dataset. Figures 3-10 above were also created for the top 5 companies by average volume looking at each centrality measures for the first and last week of the dataset.

Closeness:

Across the 5 years, NCLH remained a prominent node in terms of closeness centrality, indicating its continued influence and connectivity within the network over the years. MU also remained a key node, maintaining its position as a highly central node in the network. New entries like NVDA and BAC show that the network dynamics have changed, with these nodes becoming more central over time. Overall, the closeness centrality values have decreased, which may suggest a more decentralized or less interconnected network.

Betweenness:

NCLH appears again in 2023, indicating its increasing importance in bridging different parts of the network.cAAL and WMT are new top nodes in terms of betweenness centrality, suggesting their roles as crucial intermediaries have grown. AAPL has dropped out of the top 5, which might indicate a change in its role or influence in the network. The presence of SHOP in both timeframes highlights its continued significance over time.

PageRank:

The PageRank values in 2023 are higher compared to 2019, indicating a shift towards nodes having more influence within the network. AAPL and TSLA no longer appear in the top 5, replaced by AAL, NCLH, and WMT, suggesting a significant shift in the network's influential nodes. NCLH again shows up prominently, confirming its rising importance across different centrality measures.

Eigenvector:

AAL shows the highest eigenvector centrality in 2023, indicating it is connected to other highly influential nodes. SWN and SHOP, which were significant in 2019, have dropped out of the top 5, replaced by C, AMD, and PCG. NCLH does not appear in the top eigenvector centrality nodes for 2023, which might suggest a shift in the nodes it is connected to or a change in its overall influence.

The comparison of centrality measures between 1-10-2019 and 12-28-2023 highlights significant shifts in network dynamics. Certain nodes like NCLH and MU maintain their importance across the years, while new influential nodes such as NVDA, AAL, and WMT emerge. This suggests changes in the structure and connectivity of the network, reflecting possible changes in market conditions, company performances, and overall economic factors over the years.

c. C. Centrality Measures Over Time

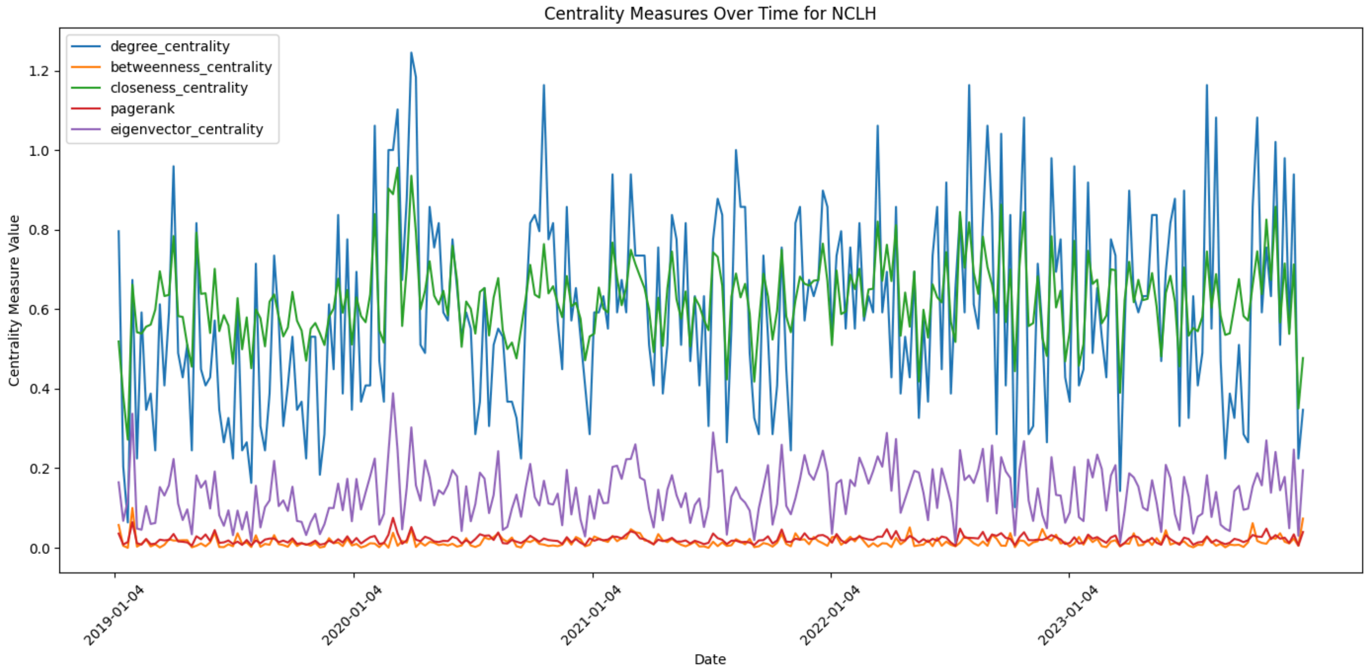


Fig. 3 Centrality Measures Over Time for NCLH

As centrality measures are used as exogenous features, we can examine their fluctuations over the five-year period. NCLH remained influential within the network. Therefore, let's analyze the graph of its centrality measures over time. Figure 3 above illustrates the evolution of the centrality measures for NCLH from 2019 to 2023, including degree centrality, betweenness centrality, closeness centrality, PageRank, and eigenvector centrality. Degree centrality, represented by the blue line, shows significant fluctuations, indicating varying levels of connectivity for NCLH within the network over time. A high peak occurs around early 2020, suggesting periods when NCLH had a higher number of direct connections with other nodes. Betweenness centrality, depicted by the orange line, remains relatively low and stable, indicating that NCLH did not frequently act as a bridge between other nodes in the network. Closeness centrality, shown in green, also fluctuates but follows a somewhat smoother trend compared to degree centrality, suggesting changing but relatively consistent overall influence within the network. PageRank, marked in red, stays consistently low throughout the period, highlighting that NCLH did not significantly dominate the network in terms of influence spread. Eigenvector centrality, represented by the purple line, shows moderate fluctuations, indicating that while NCLH's connections to highly influential nodes varied, it maintained some level of importance within the network. Overall, the graph highlights the dynamic nature of NCLH's role in the network, with variability in its direct connections, influence, and bridging role over the four-year period.

d. D. Model Results

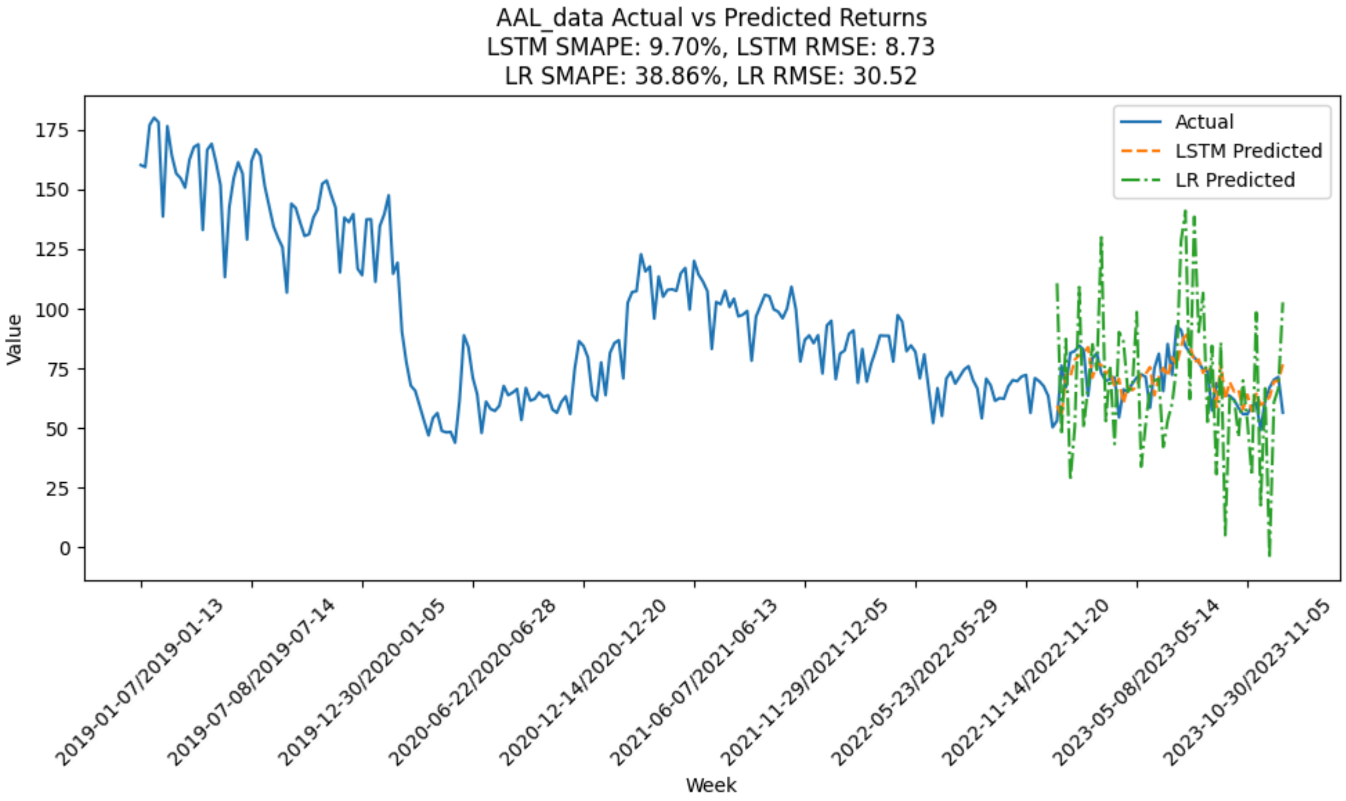
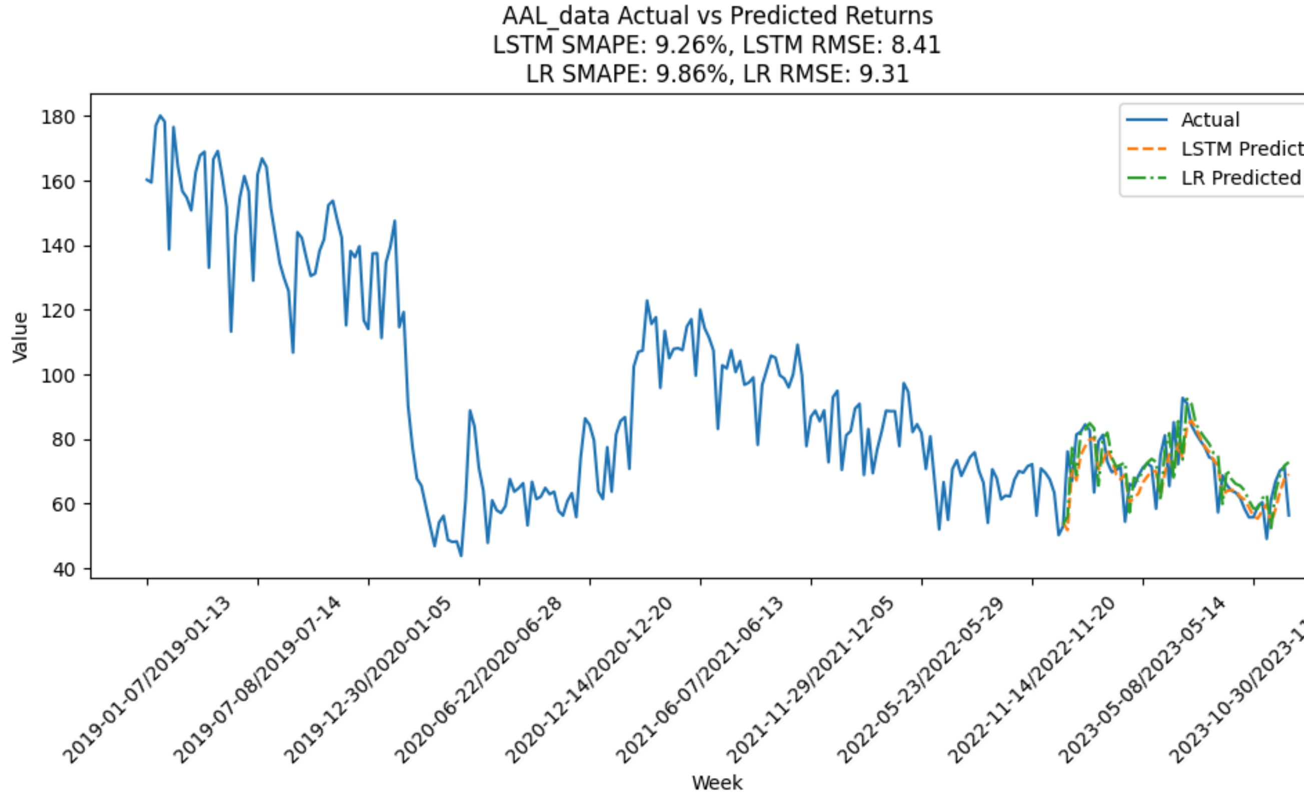


Fig. 6 Actual versus Predicted Returns for AAL with Using Centrality Measures

Figure 6 above presents the actual versus predicted stock returns using centrality measures as features for AAL. The LSTM model shows a SMAPE (Symmetric Mean Absolute Percentage Error) of 9.70% and an RMSE (Root Mean Square Error) of 8.73. The Linear Regression model, on the other hand, exhibits a much higher SMAPE of 38.86% and an RMSE of 30.52. The LSTM model significantly outperforms the Linear Regression model in this scenario. The actual returns (blue line) show a clear downward trend followed by a period of volatility. The LSTM model (orange dashed line) captures this trend and the subsequent fluctuations more accurately than the Linear Regression model (green dashed-dotted line). The higher error metrics of the Linear Regression model indicate its limited ability to leverage the centrality measures effectively for prediction, suggesting that the non-linear nature of the LSTM model is more suitable for capturing complex patterns in the data.



e.

Fig. 7 Actual vs Predicted Returns for AAL without Using Centrality Measures

Figure 7 above illustrates the actual versus predicted stock returns using only the target stock returns as features. Here, the LSTM model achieves a SMAPE of 9.26% and an RMSE of 8.41, while the Linear

Regression model achieves a SMAPE of 9.86% and an RMSE of 9.31. Both models perform similarly when using the target stock returns as features, with the LSTM model showing a slight edge in both SMAPE and RMSE. The actual returns (blue line) and the predicted returns from both models (LSTM: orange dashed line, LR: green dashed-dotted line) closely follow each other, indicating that the models can effectively predict returns when the feature set is the stock returns themselves. The relatively lower error metrics for both models in this scenario suggest that past stock returns are a strong predictor of future returns, reducing the complexity needed to capture the trends.

f. E. With and without Centrality Measures Model Comparison

The LSTM model demonstrates consistent performance across both feature sets, suggesting its flexibility and strength in capturing both linear and non-linear patterns. The Linear Regression model, while performing well with straightforward features, struggles with complex, derived features such as centrality measures, highlighting its limitations in dealing with non-linear data.

Table 9
Model Performance

	With Centrality Measures	Without Centrality Measures
LSTM sMAPE	11.35%	11.66%
LSTM RMSE	28.26	28.99
LR sMAPE	41.70%	11.55%
LR RMSE	77.08	30.77

Table 9 above aggregates the overall performance of the LSTM and Linear Regression (LR) models in predicting stock returns for all 50 companies. In order to determine the performance, overall sMAPE and RMSE values were calculated for the models with and without using centrality measures as exogenous variables.

The incorporation of centrality measures into the LSTM model yielded a sMAPE of 11.35% and an RMSE of 28.26, demonstrating a slight improvement over the LSTM model that excluded these measures, which achieved a SMAPE of 11.66% and an RMSE of 28.99. Although the decrease in sMAPE by 0.31% and the reduction in RMSE by 0.73 may seem modest, these results highlight the potential benefits of integrating centrality measures into predictive models.

Centrality measures, derived from the network structure of stock correlations, provide valuable insights into the influence and interconnectedness of stocks within the market. The observed improvements in both sMAPE and RMSE indicate that these network-based features can enhance the model's ability to capture complex market dynamics, leading to more accurate predictions. The slight decrease in error metrics suggests that centrality measures contribute positively by offering additional context that pure return data may lack. This enhancement underscores the promise of further exploring and refining the use of network analytics in financial modeling. The integration of such measures could be crucial for developing more robust predictive tools that leverage the interconnected nature of financial markets.

The Linear Regression model with centrality measures showed a significantly higher sMAPE of 41.70% and an RMSE of 77.08. This indicates a poor fit when using centrality measures, suggesting that the Linear Regression model struggles to capture the complex relationships within these features. The Linear Regression model without centrality measures achieved a sMAPE of 11.55% and an RMSE of 30.77. The performance improved dramatically when centrality measures were excluded, showing that Linear Regression works better with simpler, more direct features like past stock returns. The stark difference in performance suggests that while centrality measures provide valuable insights into stock market dynamics, their complexity may overwhelm simpler models like Linear Regression. Linear Regression is fundamentally designed to work well with straightforward, linear relationships. The intricate, non-linear interactions captured by centrality measures likely introduce noise and complexity that the model cannot adequately process, leading to poorer predictive accuracy.

This observation underscores the importance of selecting appropriate modeling techniques for different types of features. While centrality measures hold promise, their effective utilization may require more sophisticated models capable of handling complex, non-linear data structures, such as LSTM networks or SARIMAX (Seasonal AutoRegressive Integrated Moving Average with eXogenous regressors) models which is designed to capture the temporal dependencies and seasonal patterns in time series data, making them better suited for financial data.

VII. Conclusion

This project explored the enhancement of stock return predictions by leveraging network centrality measures based on maximized lagged correlations. This approach incorporated both network construction and predictive modeling, aiming to harness the value of centrality measures in forecasting stock returns.

The results indicated a nuanced impact of centrality measures on predictive performance. The LSTM model demonstrated consistent performance across both feature sets, achieving a slight improvement when centrality measures were included (sMAPE: 11.35%, RMSE: 28.26) compared to using only the target stock returns (sMAPE: 11.66%, RMSE: 28.99). This suggests that the LSTM model demonstrates potential to leverage the complexity of centrality measures effectively, showcasing its ability to enhance predictive accuracy.

On the other hand, the Linear Regression model exhibited a stark contrast in performance. When centrality measures were included, the model's error metrics increased significantly (sMAPE: 41.70%, RMSE: 77.08), indicating a poor fit. Conversely, using only the target stock returns resulted in much better performance (sMAPE: 11.55%, RMSE: 30.77). This highlights the limitations of Linear Regression in handling non-linear, complex relationships inherent in network-based features.

Overall, the analysis underscores the importance of model selection and feature engineering in financial predictive analytics. The LSTM model's ability to handle complex patterns makes it a more suitable choice for incorporating advanced features like centrality measures. However, the observed gains highlight the potential for further refinement and exploration of feature combinations to unlock even greater improvements in predictive performance.

While centrality measures offer valuable insights into the network dynamics of stock markets, their integration into predictive models requires sophisticated algorithms like LSTM to fully capitalize on their potential. This project contributes to the growing body of research on using network analytics for financial forecasting, providing a foundation for future studies to build upon and refine these techniques.

References

[1] K. Fischer and A. Palasvirta, "High Road to a Global Marketplace: The International Transmission of Stock Market Fluctuations," The Financial Review, vol. 25, pp. 371-394, 1990. doi: 10.1111/J.1540-6288.1990.TB00802.X.

[2] S. K. Stavroglou, A. A. Pantelous, H. E. Stanley, and K. M. Zuev, "Hidden interactions in financial markets," Proc. Natl. Acad. Sci. U. S. A., vol. 116, no. 22, pp. 10646-10651, May 2019. doi: 10.1073/pnas.1819449116.

[3] T. Isogai, "Dynamic correlation network analysis of financial asset returns with network clustering," Appl. Netw. Sci., vol. 2, no. 8, 2017. doi: 10.1007/s41109-017-0031-6.

[4] V. Boginski, S. Butenko, and P. Pardalos, "Mining market data: A network approach," Comput. Oper. Res., vol. 33, pp. 3171-3184, 2006. doi: 10.1016/j.cor.2005.01.027.

[5] X. Guo, H. Zhang, and T. Tian, "Development of stock correlation networks using mutual information and financial big data," PLoS ONE, vol. 13, no. 4, p. e0195941, 2018. doi: 10.1371/journal.pone.0195941.

[6] "Network Analysis of the Stock Market," Stanford University, 2015. [Online]. Available: https://snap.stanford.edu/class/cs224w-2015/projects_2015/Network_Analysis_of_the_Stock_Market.pdf

[7] M. Kim and H. Sayama, "Predicting stock market movements using network science: an information theoretic approach," Appl. Netw. Sci., vol. 2, no. 35, 2017. doi: 10.1007/s41109-017-0055-y.

[8] J. Stübinger and D. Walter, "Using Multi-Dimensional Dynamic Time Warping to Identify

Time-Varying Lead-Lag Relationships," Sensors (Basel), vol. 22, no. 18, p. 6884, Sep. 2022. doi: 10.3390/s22186884.

[9] Y. Li, T. Wang, B. Sun, et al., "Detecting the lead–lag effect in stock markets: definition, patterns, and investment strategies," Financ. Innov., vol. 8, no. 51, 2022. doi: 10.1186/s40854-022-00356-3.

[10] L. S. Junior, A. Mullokandov, and D. Y. Kenett, "Dependency Relations among International Stock Market Indices," J. Risk Financial Manag., vol. 8, pp. 227-265, 2015. doi: 10.3390/jrfm8020227.

[11] F. A. Rodrigues, "Network centrality: an introduction," arXiv, 2019. [Online]. Available: https://arxiv.org/abs/1901.07901

[12] M. Kim and H. Sayama, "Predicting stock market movements using network science: An information theoretic approach," arXiv, 2017. [Online]. Available: https://arxiv.org/abs/1705.07980.

[13] [13] D. Castilho, T. T. P. Souza, S. M. Kang, J. Gama, and A. C. P. L. F. de Carvalho, "Forecasting Financial Market Structure from Network Features using Machine Learning," arXiv, 2021. [Online]. Available: https://arxiv.org/abs/2110.11751.

[14] G. Peralta and A. Zareei, "A Network Approach to Portfolio Selection," SSRN, 2016. doi: 10.2139/ssrn.2430309.

[15] T. K. Lee, J. H. Cho, D. S. Kwon, and S. Y. Sohn, "Global stock market investment strategies based on financial network indicators using machine learning techniques," Expert Syst. Appl., vol. 95, pp. 191-198, 2018. doi: 10.1016/j.eswa.2017.11.045.

[16] X. Guo, H. Zhang, and T. Tian, "Development of stock correlation networks using mutual information and financial big data," PLoS ONE, vol. 13, no. 4, p. e0195941, 2018. doi: 10.1371/journal.pone.0195941.

[17] S. Bennett, M. Cucuringu, and G. Reinert, "Lead–lag detection and network clustering for multivariate time series with an application to the US equity market," Mach. Learn., vol. 111, pp. 4497-4538, 2022. doi: 10.1007/s10994-022-06250-4.

[18] O. Surakhi, M. A. Zaidan, P. L. Fung, N. H. Motlagh, S. Serhan, M. Alkhanafseh, R. Ghoniem, and T. Hussein, "Time-Lag Selection for Time-Series Forecasting Using Neural Network and Heuristic

Algorithm," Electronics, vol. 10, no. 20, p. 2518, 2021. doi: 10.3390/electronics10202518.

[19] L. Nevasalmi, "Forecasting Multinomial Stock Returns Using Machine Learning Methods," Econometrics: Econometric & Statistical Methods - Special Topics eJournal, 2020. doi: 10.2139/ssrn.3630222.

[20] T. G. Bali, A. Goyal, D. Huang, F. Jiang, and Q. Wen, "Predicting Corporate Bond Returns: Merton Meets Machine Learning," Georgetown McDonough School of Business Research Paper No. 3686164, Swiss Finance Institute Research Paper No. 20-110, 2020. [Online]. Available: https://ssrn.com/abstract=3686164.

[21] J. Brownlee, *"Time Series Prediction with LSTM Recurrent Neural Networks in Python with Keras," Machine Learning Mastery,* 2023. [Online]. Available: https://machinelearningmastery.com/time-series-prediction-lstm-recurrent-neural-networks-python-keras/. [Accessed: 16-Jun-2024].