https://doi.org/10.26219/heal.aueb.6624

ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ
ATHENS UNIVERSITY OF ECONOMICS AND BUSINESS
ΣΧΟΛΗ ΕΠΙΣΤΗΜΩΝ & ΤΕΧΝΟΛΟΓΙΑΣ ΤΗΣ ΠΛΗΡΟΦΟΡΙΑΣ
SCHOOL OF INFORMATION SCIENCES & TECHNOLOGY
ΜΕΤΑΠΤΥΧΙΑΚΟ ΣΤΑΤΙΣΤΙΚΗ
MSc IN STATISTICS

SCHOOL OF INFORMATION SCIENCES & TECHNOLOGY

DEPARTMENT OF STATISTICS
POSTGRADUATE PROGRAM

Quantifying Credit Risk in Complex Financial Networks

By
Dimitrios Liordos

M.Sc. Thesis
Submitted to the Department of Statistics
of the Athens University of Economics and Business
in partial fulfilment of the requirements for
the degree of Master of Science in Statistics

Athens, Greece
October 2023

https://doi.org/10.26219/heal.aueb.6624

ΣΧΟΛΗ ΕΠΙΣΤΗΜΩΝ & ΤΕΧΝΟΛΟΓΙΑΣ ΤΗΣ ΠΛΗΡΟΦΟΡΙΑΣ

ΤΜΗΜΑ ΣΤΑΤΙΣΤΙΚΗΣ
ΜΕΤΑΠΤΥΧΙΑΚΟ ΠΡΟΓΡΑΜΜΑ

Ποσοτικοποίηση Πιστωτικού Κινδύνου σε Πολύπλοκα Χρηματοοικονομικά Δίκτυα

Δημήτριος Λιόρδος

Διατρηβή
Που υποβλήθηκε στο Τμήμα Στατιστικής
του Οικονομικού Πανεπιστημίου Αθηνών
ως μέρος των απαιτήσεων για την απόκτηση
Μεταπτυχιακού Διπλώματος Ειδίκευσης στη Στατιστική

Αθήνα
Οκτώβριος 2023

https://doi.org/10.26219/heal.aueb.6624

Abstract

Quantifying Credit Risk in Complex Financial Networks

by Dimitrios Liordos

October 2023

In today's complex financial landscape, our thesis addresses a critical problem: the need for enhanced credit risk assessment methodologies. The modern financial system is characterized by complex interconnections, and understanding the implications of these networks is paramount for effective credit risk management. This thesis aims to contribute to this understanding by exploring the role of networks in credit risk assessment. We first establish the foundational knowledge, introducing key concepts of network analysis, such as network representations and metrics. This background knowledge is crucial for navigating the complexities of financial networks. Next, we present a literature review regarding systemic risk and network analysis, community detection and shareholders networks. The research methodology, involves the collection and analysis of network data. We examine bipartite networks, projecting into shareholder-shareholder and company-company networks, and employ network metrics to derive insights into the dynamics of shareholders and companies. Moreover, we introduce two novel network metrics: the 'Company Score' and 'Distance from Default', where we define them in detail and illustrate their practical relevance. To assess their performance, we constructed three logistic regression models. These models include one with traditional network variables, one integrating the company score, and one incorporating the distance from default as predictors, with the objective to evaluate the predictive capabilities of these two novel network metrics in comparison to traditional ones. Lastly, our research highlights the potential of network analysis as a valuable tool in credit risk assessment. It sheds light on hidden patterns within financial networks, offering a new perspective for risk management. This work contributes to the broader field of financial analytics and risk assessment.

 https://doi.org/10.26219/heal.aueb.6624

Περίληψη

Ποσοτικοποίηση Πιστωτικού Κινδύνου σε Πολύπλοκα Χρηματοοικονομικά Δίκτυα

Δημήτριος Λιόρδος

Οκτώβριος 2023

Στο σημερινό περίπλοκο χρηματοοικονομικό τοπίο, η διατριβή αντιμετωπίζει ένα κρίσιμο πρόβλημα: την ανάγκη για βελτιωμένες μεθοδολογίες αξιολόγησης πιστωτικού κινδύνου. Το σύγχρονο χρηματοοικονομικό σύστημα χαρακτηρίζεται από πολύπλοκες διασυνδέσεις και η κατανόηση των επιπτώσεων αυτών των δικτύων είναι πρωταρχικής σημασίας για την αποτελεσματική διαχείριση πιστωτικού κινδύνου. Η παρούσα διατριβή στοχεύει να συμβάλει σε αυτή την κατανόηση διερευνώντας τον ρόλο των δικτύων στην αξιολόγηση του πιστωτικού κινδύνου. Αρχικά καθιερώνουμε τις θεμελιώδεις γνώσεις, εισάγοντας βασικές έννοιες της ανάλυσης δικτύου, όπως αναπαραστάσεις και μέτρα δικτύου. Αυτές οι βασικές γνώσεις είναι ζωτικής σημασίας για την ανάλυση πολύπλοκων χρηματοοικονομικών δικτύων. Στη συνέχεια, παρουσιάζουμε μια βιβλιογραφική ανασκόπηση σχετικά με τον συστημικό κίνδυνο και την ανάλυση δικτύου, το 'community detection' και 'shareholders networks'. Η μεθοδολογία της έρευνας μας, περιλαμβάνει τη συλλογή και ανάλυση δεδομένων δικτύου. Εξετάζουμε διμερή δίκτυα, προβάλλοντας σε δίκτυα μετόχων-μετόχων και εταιρειών-εταιρειών και χρησιμοποιούμε μέτρα δικτύου για να αντλήσουμε γνώσεις σχετικά με τη δυναμική των μετόχων και των εταιρειών. Επιπλέον, εισάγουμε δύο νέα μέτρα δικτύου: το 'Company Score' και το 'Distance from Default', όπου τα ορίζουμε λεπτομερώς και παρουσιάζουμε την πρακτική τους σημασία. Για να αξιολογήσουμε την απόδοσή τους, κατασκευάσαμε τρία μοντέλα λογιστικής παλινδρόμησης. Αυτά τα μοντέλα περιλαμβάνουν ένα με τις συνήθεις μεταβλητές δικτύου, ένα που ενσωματώνει το 'Company Score' και ένα που ενσωματώνει το 'Distance from Default' ως ανεξάρτητες μεταβλητές, με στόχο την αξιολόγηση των προγνωστικών ικανοτήτων αυτών των δύο νέων μετρών δικτύου σε σύγκριση με αυτές που υπάρχουν στην βιβλιογραφία. Τέλος, η έρευνά αυτή δίνει έμφαση στις δυνατότητες της ανάλυσης δικτύων ως πολύτιμου εργαλείου για την αξιολόγηση πιστωτικού κινδύνου και στην εύρεση κρυφών μοτίβων μέσα στα χρηματοοικονομικά δίκτυα, προσφέροντας μια νέα προοπτική για τη διαχείριση κινδύνου. Αυτή η εργασία συμβάλλει στο ευρύτερο πεδίο της χρηματοοικονομικής ανάλυσης και της αξιολόγησης κινδύνου.

https://doi.org/10.26219/heal.aueb.6624

Contents

Abstract **i**

List of Tables **v**

List of Figures **vi**

1 Introduction **1**
1.1 Problem Statement 1
1.2 Overview of Thesis Structure 2

2 Network Analysis Preliminaries **3**
2.1 Representations of Networks 3
2.2 Network Metrics 6
2.2.1 Degree Centrality 7
2.2.2 Eigenvector Centrality 7
2.2.3 Katz Centrality 9
2.2.4 PageRank 10
2.2.5 Closeness Centrality 12
2.2.6 Betweenness Centrality 12
2.2.7 Groups of Vertices 13
2.2.8 Clustering Coefficient 14
2.3 An Illustrative Example 15

3 Literature Review **17**
3.1 Systemic Risk and Network Analysis 17
3.2 Community Detection in Networks 19
3.3 Shareholders Networks 20

4 Understanding the Network: Data and Analysis **22**
4.1 Data Overview . . . 22
4.2 Bipartite Network Analysis . . . 23
4.2.1 Degree Centrality . . . 23
4.2.2 Betweenness Centrality . . . 25
4.3 Unipartite Network Analysis . . . 27
4.3.1 Projected Networks . . . 27
4.3.2 Shareholder-Shareholder Network Analysis . . . 28
4.3.3 Company-Company Network Analysis . . . 34

5 Contributions and Results **40**
5.1 Novel Network Metrics . . . 40
5.1.1 Company Score . . . 41
5.1.2 Distance from Default . . . 42
5.2 Logistic Regression Models . . . 46
5.2.1 Data Selection . . . 46
5.2.2 Logistic Regression with Network Variables . . . 48
5.2.3 Logistic Regression with Company Score . . . 49
5.2.4 Logistic Regression with Distance from Default . . . 51
5.3 Results . . . 52

6 Conclusion **54**
6.1 Summary of Findings . . . 54
6.2 Future Work . . . 55

Bibliography **57**

 https://doi.org/10.26219/heal.aueb.6624

List of Tables

2.1 Different network metrics for each vertex 15

4.1 Top five shareholders for each network centrality measure. 34
4.2 Top five companies for each network centrality measure. 39

5.1 Shareholder-Company relationships 42
5.2 Summary of the model with the network variables as predictors . . . 48
5.3 Different measures for the evaluation of logistic regression with network variables 49
5.4 Summary of the model with the company score as predictor 50
5.5 Different measures for the evaluation of logistic regression with the company score 50
5.6 Summary of the model with the distance from default as a predictor . 51
5.7 Different measures for the evaluation of logistic regression with the distance from default 51
5.8 Summary of performance metrics for the three models in each quarter 53

https://doi.org/10.26219/heal.aueb.6624

List of Figures

2.1 The left one represents an undirected network and the right one a directed one. 4
2.2 A bipartite network with 4 vertices on each set, one representing the shareholders and the other the companies. 6
2.3 An undirected network with six vertices 16

4.1 Histograms of the degrees for the shareholders and the companies respectively. 24
4.2 Histograms of the betweenness centralities for the shareholders and the companies respectively. 26
4.3 Bipartite network on the left and its two projections on the right. . . . 28
4.4 Shareholder-Shareholder network 29
4.5 Degree centrality histogram of the S-S network. 29
4.6 Betweenness centrality histogram of the S-S network. 30
4.7 Closeness centrality histogram of the S-S network. 31
4.8 Eigenvector centrality histogram of the S-S network. 32
4.9 Clustering coefficient histogram of the S-S network. 33
4.10 Company-Company network. 34
4.11 Degree centrality histogram of the C-C network. 35
4.12 Betweenness centrality histogram of the C-C network. 36
4.13 Closeness centrality histogram of the C-C network. 36
4.14 Eigenvector centrality histogram of the C-C network. 37
4.15 Clustering coefficient histogram of the C-C network. 38
4.16 Distribution of shortest path length in the C-C network. 39

5.1 A bipartite graph on the left and its projection on the right. 45
5.2 Plot of the number of default event versus time. 47

https://doi.org/10.26219/heal.aueb.6624

Chapter 1

Introduction

In today's complex financial landscape, this thesis embarks on a comprehensive exploration of networks and their implications for credit risk assessment. Our focus is to uncover hidden patterns and valuable insights within financial networks that can be applied to enhance our ability to predict and manage credit risk.

1.1 Problem Statement

The research problem addressed in this thesis revolves around the exploration of novel predictive metrics derived from complex financial networks to enhance the accuracy of company default risk prediction models. By delving into the complex structures and dynamics of financial networks, our aim is to develop innovative metrics that can effectively complement traditional predictors used in credit risk assessment. This problem is of paramount importance due to its potential to revolutionize predictive modeling in credit risk assessment. Traditional methodologies, while effective to some extent, often overlook valuable insights embedded within financial networks. By harnessing these insights, we can improve the predictive power of credit risk assessment models.

The potential contributions of this research to the field are multifaceted. Firstly, it offers a novel approach to credit risk assessment by integrating network-based metrics into predictive models. This approach has the potential to enhance the accuracy and reliability of default risk predictions, leading to more effective risk management strategies. Additionally, the exploration of complex financial networks contributes to a deeper understanding of financial systems and their interdependencies, shedding light on systemic risks and vulnerabilities. Overall, this research opens up new av-

https://doi.org/10.26219/heal.aueb.6624

enues for advancing predictive modeling techniques and enriching our understanding of financial networks' role in credit risk assessment.

1.2 Overview of Thesis Structure

We begin our journey in chapter 2, where we lay the groundwork by introducing key concepts such as fundamental representations of networks, network metrics, and illustrate these concepts with a tangible example. These fundamental tools will be essential as we navigate through the complexities of financial networks.

Chapter 3, guiding us through the vast area of existing knowledge. We explore systemic risk and network analysis, delving into its definitions within financial systems. We traverse the landscape of community detection, a crucial aspect of network analysis, and embark into shareholders networks.

Our analysis, starts in chapter 4 where we delve into the complexities of network data and explore bipartite networks. We transition into the projected networks of the bipartite case, namely shareholder-shareholder and company-company networks. There we use the traditional network metrics that we present in chapter 2 and find interesting results for the shareholders and the companies.

In chapter 5, we introduce two novel network metrics, the 'Company Score' and 'Distance from Default'. We define this metrics in detail and provide practical examples to illustrate their relevance. Leveraging these metrics, we construct three logistic regression models aimed at predicting company defaults. The first model integrates the traditional network variables from Chapter 4, while the second and third models incorporate the newly introduced company score and distance from default metrics, respectively. Our primary objective is to evaluate the predictive efficacy of these novel network metrics relative to conventional ones.

As our exploration finishes, in chapter 6 we synthesize the wealth of insights gained throughout our thesis, offering a comprehensive summary of findings and results. Additionally, we look into the future for possibilities for further research and the continued evolution of network-based financial analytics.

https://doi.org/10.26219/heal.aueb.6624

Chapter 2

Network Analysis Preliminaries

In this chapter, we lay the groundwork by introducing essential theoretical tools derived from graph theory. This branch of mathematics is dedicated to the study of networks, offering a wealth of results and insights. While we present only a fraction of graph theory's vast expanse, we focus on those concepts most relevant for our analysis. Furthermore, we introduce different measures and metrics that quantify the network structure, enabling us to gain a deeper understanding of the network's properties and its topology. These measures were initially developed within the field of social network analysis, but now have become widely spread across numerous disciplines. Our exploration is built upon the foundational insights provided by Newman [1].

2.1 Representations of Networks

The first and most important thing that we have to mention is what is a graph (also called a network). Formally, a graph $G = (V, E)$ is a mathematical structure consisting of a set V of vertices (also commonly called nodes) and a set E of edges (also commonly called links), where elements of E are unordered pairs u, v of distinct vertices $u, v \in V$. Now, that we know what a network is, we need a way to represent it mathematically. There exist various mathematical representations for networks, each offering unique insights into their structure and connectivity.

We can consider two networks, an undirected where the direction of the edges doesn't matter, and a directed one where the direction of the edges does matter. Two examples of these networks can be seen in Figure 2.1. To fully describe the entire network, we need two pieces of information: the number of vertices and a list of all the edges

https://doi.org/10.26219/heal.aueb.6624

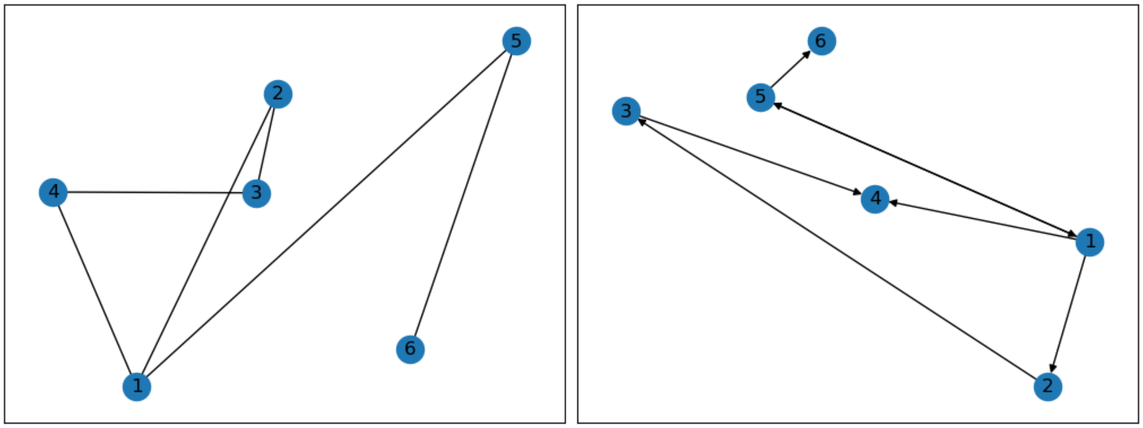

Figure 2.1: The left one represents an undirected network and the right one a directed one.

present in the network. We denote an edge between vertices i and j as (i, j). In the case of the left network depicted in Figure 2.1, we have 6 vertices, and the edges are $(1, 2)$, $(1, 5)$, $(2, 3)$, $(3, 4)$, $(5, 6)$ and $(1, 4)$. On the other hand, regarding the right network in the Figure 2.1, the direction of the edges is important. For example, the edge $(5, 6)$ it means that we go from the vertex 5 to 6 and not the other way as we did in the undirected network.

A better representation of a network is the adjacency matrix. The adjacency matrix $\mathbf{A}$ of an undirected network is the matrix with elements A_{ij} such that:

$$A_{ij} = \begin{cases} 1 & \text{if } (i, j) \in E, \\ 0 & \text{otherwise} \end{cases}$$

For example, the adjacency matrix of the left network in Figure 2.1 is:

$$A_{ij} = \begin{pmatrix} 0 & 1 & 0 & 1 & 1 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 \end{pmatrix}$$

We have to mention that for an undirected network with no self-edges, such as the one in our example, the diagonal matrix elements are all zero and the matrix is symmetric.

https://doi.org/10.26219/heal.aueb.6624

On the other hand, the adjacency matrix of a directed network has matrix elements:

$$A_{ij} = \begin{cases} 1 & \text{if there is and edge from } j \text{ to } i, \\ 0 & \text{otherwise} \end{cases}$$

The elements of diagonal in this matrix will be again zero if we don't have self-edges but the matrix will not be symmetric.

From the above, we can see that the adjacency matrix takes only values 0 and 1. This is not always the case. Weighted networks, play a vital role in representing complex systems where connections between vertices carry different degrees of strength. In these networks, each edge is assigned a numerical weight that quantifies the intensity or importance of the relationship between nodes. These networks can be represented by an adjacency matrix, but with element values equal to the weights of the corresponding connections.

Lastly, we are going to discuss bipartite networks, which provide a foundational framework where complex relationships between two distinct sets of entities can be effectively represented and analyzed. This graph structure divides nodes into two separate and non-intersecting sets, with connections exclusively established between nodes belonging to different groups. For instance, in the context of a shareholder-company network, one set of nodes corresponds to shareholders, while the other set represents companies. The presence of an edge between a shareholder and a company indicates that the shareholder has invested in this company. A visual representation of this network can be seen in Figure 2.2, where S represents the shareholders and C the companies. Moreover, the equivalent of an adjacency matrix for a bipartite network is a rectangular matrix called an incidence matrix. If n is the number of shareholders in the network and g is the number of companies, then the incidence matrix $\mathbf{B}$ is a $g \times n$ matrix having elements B_{ij} such that:

$$B_{ij} = \begin{cases} 1 & \text{if vertex } j \text{ belongs to group } i, \\ 0 & \text{otherwise} \end{cases}$$

So for our example the incidence matrix will be:

$$B_{ij} = \begin{pmatrix} 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix}$$

https://doi.org/10.26219/heal.aueb.6624

Furthermore, one of the valuable insights that can be derived from bipartite graphs is the concept of projections. A projection involves transforming a bipartite network into a unipartite network by aggregating connections between nodes of one group based on their shared interactions with nodes from the other group. In the shareholder-company network example, a shareholder-shareholder projection could reveal common investment portfolios, enabling the identification of shareholder communities with similar investment behavior. In general, projections offer a deeper understanding of the underlying relationships and facilitate the application of traditional network analysis techniques to extract meaningful patterns and insights.

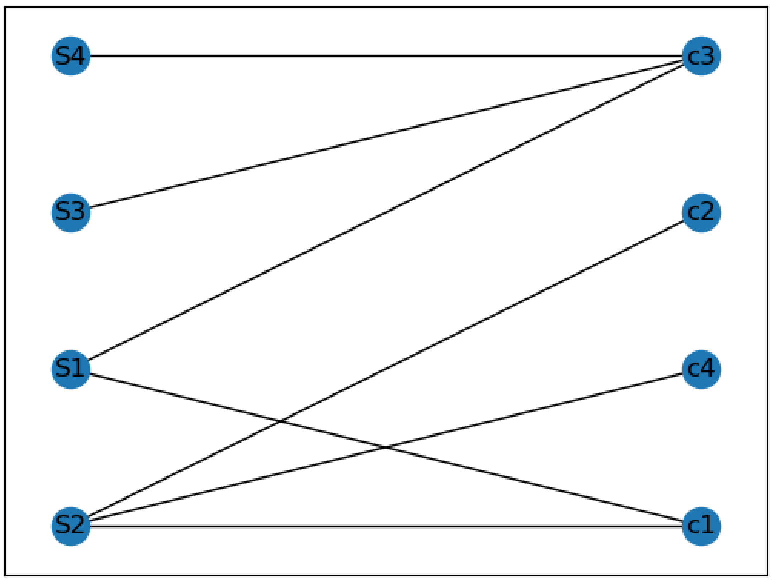

Figure 2.2: A bipartite network with 4 vertices on each set, one representing the shareholders and the other the companies.

2.2 Network Metrics

Now that we have a better understanding of what a network is and how we can represent it mathematically, we can use this knowledge and move to the introduction of different metrics that will help us gain a deeper understanding of the network's properties and topology.

https://doi.org/10.26219/heal.aueb.6624

2.2.1 Degree Centrality

The first and the simplest centrality measure that we will introduce is the degree or degree centrality (degree divided by the total number of vertices) of a vertex, which is the number of edges connected to it. For an undirected graph of n vertices, the degree can be written as:

$$k_i = \sum_{j=1}^{n} A_{ij}, \quad \text{where A is the adjacency matrix} \tag{2.1}$$

On the other hand, in a directed network each vertex has two degrees. The in-degree is the number of in-going edges connected to a vertex and the out-degree is the number of outgoing edges. Similarly, as before, we can write them as:

$$k_i^{in} = \sum_{j=1}^{n} A_{ij} \quad k_j^{out} = \sum_{i=1}^{n} A_{ij} \tag{2.2}$$

Also, we have to remember that the adjacency matrix of a directed network has element $A_{ij} = 1$ if there is an edge from j to i and not the other way around.

2.2.2 Eigenvector Centrality

Next, we will introduce the eigenvector centrality which serves as a natural extension of the degree centrality. While degree centrality assigns equal "centrality points" to a vertex based on the number of edges that it has and assumes that all the neighbors are the same importance, eigenvector centrality acknowledges that not all neighbors carry the same level of information. In many circumstances, a vertex's significance within a network is increased by having connections to other vertices that are themselves important. This forms the core idea behind eigenvector centrality. Instead of simply granting one point per neighbor, eigenvector centrality assigns each vertex a score that is proportional to the sum of the scores of its neighbors.

To begin, we can establish an initial estimate for the centrality, denoted as x_i of each vertex i. As a starting point, we can set $x_i = 1$ for all vertices. Although this assignment may not provide a meaningful measure of centrality, we can utilize it as a foundation to compute an improved measure x_i'. We can define it to be the sum of the centralities of i's neighbors:

$$x_i' = \sum_j A_{ij} x_j$$

where A_{ij} is an element of the adjacency matrix. We can also write this in matrix notation as $\mathbf{x}^{'} = \mathbf{Ax}$, where x is the vector with elements x_i. Repeating this process to make better estimates, we have after t steps a vector of centralities $x(t)$ given by:

$$x(t) = A^t x(0)$$

We can write $x(0)$ as a linear combination of the eigenvectors v_i of the adjacency matrix:

$$x(0) = \textstyle\sum_i c_i v_i,$$

for some appropriate choice of constants c_i. Then we have that:

$$x(t) = A^t \sum_i c_i v_i = \sum_i c_i \kappa_i^t v_i = \kappa_1^t \sum_i c_i \left[\frac{\kappa_i}{\kappa_1}\right]^t v_i$$

$$= \kappa_1^t \left(c_1 \frac{\kappa^t}{\kappa_1^t} + c_2 \frac{\kappa_2^t}{\kappa_1^t} + c_3 \frac{\kappa_3^t}{\kappa_1^t} + ... \right)$$

$$x(t) \to \kappa_1^t c_1 v_1, \quad \text{since } \frac{\kappa_i^t}{\kappa_1^t} \to \infty \text{ as } t \to \infty$$

where the κ_i are the eigenvalues of A and κ_1 is the largest of them. All terms in the sum other than the first decay exponentially as t becomes large, and in the limit, we get $x(t) \to c_1 \kappa_1^t v_1$. Equivalently, we could say that the centrality x satisfies:

$$\mathbf{Ax} = \kappa_1 \mathbf{x}$$

This is the eigenvector centrality, which was first proposed by Bonacich in 1987 [2]. We can write this as:

$$x_i = \kappa_1^{-1} \sum_j A_{ij} x_j, \tag{2.3}$$

this allows us to have large centrality values from two factors: a vertex having a significant number of neighbors and/or being connected to highly influential neighbors.

Eigenvector centrality can be calculated for both undirected and directed networks in theory. However, it is most effective when applied to undirected networks. On the other hand, dealing with directed networks introduces additional complexities. One significant difference is that a directed network's adjacency matrix is generally non symmetric, leading to the existence of two sets of eigenvectors (left and right). In the

https://doi.org/10.26219/heal.aueb.6624

majority of cases, the right eigenvector is the appropriate choice because in directed networks centrality is typically determined by other vertices pointing towards you, rather than you pointing to others. At the same time, this creates another problem. If we have a vertex that has only outgoing edges and no incoming ones, the centrality will be zero because there will be no terms in the sum, in Eq 2.3 This leads us, to the next measure.

2.2.3 Katz Centrality

The next measure that we will introduce is a solution to the problems that we encounter previously with the eigenvector centrality. We give each vertex a small amount of centrality, regardless of its position in the network or the centrality of its neighbors. With that we define the Katz centrality as:

$$x_i = \alpha \sum_j A_{ij} x_j + \beta \tag{2.4}$$

where α and β are positive constants. As we can see the first term of the Eq 2.4 is almost the same as the one in the eigenvector centrality and the second term is the constant extra term that all vertices receive. This additional term ensures that even vertices with zero in-degree receive a baseline centrality value equal to β. Once a vertex has non-zero centrality value due to this baseline, the vertices it points to gain some advantage from being pointed to. Consequently, any vertex that is pointed to by numerous other vertices will have a higher centrality, although vertices pointed to by others with high centrality will still receive a further boost in their centrality score.

Furthermore, we can write the Eq 2.4 in matrix notation as follows:

$$\mathbf{x} = \alpha \mathbf{A} \mathbf{x} + \beta \mathbf{I}$$

$$\mathbf{x} = \beta (\mathbf{I} - \alpha \mathbf{A})^{-1} \mathbf{I}$$

where $\mathbf{I}$ is the vector $(1, 1, 1, ...)$. We normally don't care about the absolute magnitude of the centrality, only about which vertices have high or low centrality value, thus for convenience we can set $\beta = 1$, giving:

$$\mathbf{x} = (\mathbf{I} - \alpha \mathbf{A})^{-1} \mathbf{I} \tag{2.5}$$

The above centrality measure was first proposed by Katz [3] in 1953 and we will refer to it as the Katz centrality.

https://doi.org/10.26219/heal.aueb.6624

The Katz centrality distinguishes itself from eigenvector centrality through the introduction of a parameter α, responsible for determining the trade-off between the eigenvector term and the constant term in Eq 2.4. Before calculating the Katz centrality, the constant term's value must be carefully selected. When $\alpha \to 0$, only the constant term remains in Eq.1.4, resulting in all vertices having the same centrality β (with a value set to 1). As α increase from zero, the centralities of vertices also increase until a point where they start to diverge. This critical point occurs when the expression $(\mathbf{I} - \alpha\mathbf{A})^{-1}$ diverges, specifically when $\det(\mathbf{I} - \alpha\mathbf{A})$ passes through zero. This condition can be rephrased as:

$$\det(\mathbf{A} - \alpha^{-1}\mathbf{I})$$

Observing this expression, we can recognize it as the characteristic equation with roots α^{-1} being equivalent to the eigenvalues of the adjacency matrix. As α increases, the determinant reaches zero for the first time when α^{-1} equals to κ_1, representing the largest eigenvalue of A. To ensure convergence of the centrality expression, it is advisable to choose a value for α lower than this critical point.

However, determining the precise value for a beyond this point lacks clear guidelines. Many researchers have opted for values close to the maximum of $1/\kappa_1$, which heavily emphasizes the eigenvector term while minimizing the impact of the constant term. This approach yields centrality values that are closely with the ones from the eigenvector centrality but assigns small non-zero values to vertices outside the strongly connected components.

We introduce the Katz centrality as a solution to address the challenges posed by ordinary eigenvector centrality when dealing with directed networks. However, it is worth noting that Katz centrality can also be applied to undirected networks without any fundamental constraints, and there are instances where this adaption proves advantageous.

2.2.4 PageRank

The previous section introduced a characteristic of the Katz centrality that may not always be desirable. When a vertex with high Katz centrality points to multiple others, all of those other vertices also receive high centrality values. This behavior has been subject to debate, as in many cases, it is less significant if a vertex is just one among many that are pointed to. The centrality gained from being connected to a prestigious vertex gets diluted when shared with numerous other vertices.

https://doi.org/10.26219/heal.aueb.6624

For example, consider the famous Yahoo! web directory, which contains links to millions of web pages, including mine. While Yahoo! is an important website and would have high centrality according to any reasonable measure, it may not necessarily make my page very important by association. The high centrality of Yahoo! gets distributed among millions of pages, including mine, reducing its contribution to my page's centrality.

To address this issue, a variation of the Katz centrality can be defined, where the centrality a vertex receives from its network neighbors is proportional to their centrality divides by their out-degree. With this approach, vertices that point to many others pass on only a small amount of centrality to each of those others, even if their own centrality is high. Mathematically, this centrality measure can be defined as:

$$x_i = \alpha \sum_j A_{ij} \frac{x_j}{k_j^{out}} + \beta \tag{2.6}$$

This gives problems if there are vertices in the network with out-degree equal to zero. However, this is easily fixed because it is clear that vertices with no out-going edges should contribute zero to the centrality of any other vertex. This can be achieved by setting $k_j^{out} = 1$ for all such vertices.

Again, we can write this in matrix terms as:

$$\mathbf{x} = \alpha \mathbf{A} \mathbf{D}^{-1} \mathbf{x} + \beta \mathbf{I},$$

where $\mathbf{I}$ is the vector $(1, 1, ...)$ and D the diagonal matrix with elements $D_{jj} = \max(k_j^{out}, 1)$. Rearranging, we have $\mathbf{x} = \beta(\mathbf{I} - \alpha \mathbf{A} \mathbf{D}^{-1})^{-1} \mathbf{I}$, as before β plays the role only of an overall multiplier for the centrality, so we can set it equal to one and have:

$$\mathbf{x} = \mathbf{D}(\mathbf{D} - \alpha \mathbf{A})^{-1} \mathbf{I}$$

This centrality measure is commonly known as PageRank, which is the trade name given it by the Google web search corporation, which uses it as a central part of their web ranking technology [4]. PageRank proves effective on the Web because when your page has links from important pages on other websites, it suggests that your page might also be important. However, the crucial aspect of dividing by the out-degrees of pages ensures that pages merely pointing to a vast number of other pages do not pass a significant amount of centrality to each of them. As a result, networks like Yahoo! do not exert disproportionate influence on the rankings, making the centrality distribution more balanced and fairer.

https://doi.org/10.26219/heal.aueb.6624

2.2.5 Closeness Centrality

The next measure that we will introduce offers a different way of measuring centrality. It quantifies the average distance from a vertex to all other vertices in the network. Let's consider d_{ij} as the length of the geodesic path (shortest path) from vertex i to j, where the length indicates the number of edges along the path. In this context, the mean geodesic distance from vertex i to all other vertices j in the network can be expressed as follows:

$$l_i = \frac{1}{n} \sum_j d_{ij}$$

The value of this quantity is lower for vertices that, on average, have a short geodesic distance to other vertices. Such vertices are likely to have improved access to information from other vertices or possess more direct influence on them. Unlike the other centrality measures discussed in this chapter, the mean distance l_i does not follow the same principle. It assigns low values to more central vertices and high values to less central ones, which contrasts with our other measures. Due to this discrepancy, in the social network's literature, researchers often work with the inverse of l_i. This inverse measure is known as the closeness centrality:

$$C_i = \frac{1}{l_i} = \frac{n}{\sum_j d_{ij}} \tag{2.7}$$

Closeness centrality is a widely employed and intuitive centrality measure in social and network studies. However, it does have certain drawbacks. One concern is that its values typically cover a relatively narrow range, which makes it difficult to distinguish between central and less central vertices. There are better solutions to address the problems that this measure has, but because they hardly ever used in practice, we will not mention them.

2.2.6 Betweenness Centrality

Another important measure that we will mention is the betweenness centrality, which measures the significance of a vertex in a network based on its position between other vertices. Initially it was introduced by Freeman [5] in 1977. To understand betweenness centrality, we can consider a scenario where something, like messages or data packets, flows between vertices through the shortest paths in the network. If we wait for a considerable time, the number of messages passing through each vertex on their way to their destination will be proportional to the number of geodesic paths the vertex lies on. This number of geodesic paths represents the betweenness centrality

https://doi.org/10.26219/heal.aueb.6624

of the vertex. Vertices with high betweenness centrality play a critical role in controlling information exchange between others. Their removal from the network can severely disrupt communications between other vertices, emphasizing their influence. While real-world situations may not always adhere to the assumptions, betweenness centrality still serves as an approximate guide to understanding the flow of information and the influence of vertices in a network.

Mathematically, we can express the betweenness centrality for a general network by defining n_{st}^i to be the number of geodesic paths from s to t that pass through i, and g_{st} to be the total number of geodesic paths from s to t. Then the betweenness centrality of vertex i is:

$$x_i = \sum_{s \neq t \neq i \in N} \frac{n_{st}^i}{g_{st}} \tag{2.8}$$

Betweenness centrality extends its applicability to directed networks as well. In directed networks, the shortest paths between two vertices can vary depending on the direction of travel. For instance, the shortest path from vertex A to B may differ from the shortest path from B to A, and in some cases, there might be a path in one direction while no path exists in the opposite direction. Therefore, it becomes crucial to explicitly consider the path counts in both directions for each vertex pair in a directed network. Fortunately, the definition of betweenness centrality, as given in Eq 2.8, already accounts for this aspect.

2.2.7 Groups of Vertices

In this section, we will introduce some different concepts compared to the ones that we have discussed so far. Besides looking at each vertex separately and try to calculate different metrics for each one, we can divide the network into groups or communities and study the local structure of the network.

Cliques, plexes, and cores are interconnected concepts that provide valuable insights into the structural organization of complex networks. A clique is a subset of vertices of a graph such that every two distinct vertices in the clique are adjacent, meaning that every vertex is directly connected to every other vertex forming a community. A k-plex of size n is a maximal subset of n vertices within a network such that each vertex is connected to at least n-k of the others. If $k = 1$, we recover the definition of an ordinary clique. Moreover, a k-core is a maximal subset of vertices such that each is connected to at least k others in the subset. We can prove that a k-core of n vertices is also an (n-k)-plex. Also, unlike k-plexes and cliques, k-cores cannot

https://doi.org/10.26219/heal.aueb.6624

overlap, since by their definition two k-cores that shared one or more vertices would just form a single larger k-core.

In order to gain a deeper understanding of the above, we can consider a social network where individuals represent vertices and friendships represent edges. A clique within this network could be a group of friends who all know each other. A plex might involve a set of individuals who frequently interact with each other, even if not all of them are directly connected. In this scenario, a core would be a group of individuals who maintain the most substantial connections, possibly driving the overall cohesion of the larger network. Understanding these structures helps us comprehend how information or influence spreads within the network.

2.2.8 Clustering Coefficient

Building upon the previous concepts that we introduced, we can move to the exploration of transitivity and clustering coefficient. As cliques, plexes, and cores explain the interconnectedness of nodes within a network, transitivity and clustering coefficient provide quantitative measures for finding which nodes tend to form clusters or groups.

Transitivity evaluates the likelihood that if node A is connected to node B, and node B is connected to node C, then node A and node C are also connected. In essence, transitivity captures the tendency for triangles or closed loops to emerge within networks. Furthermore, the global clustering coefficient, closely related to transitivity, is a measure of the degree to which nodes in a graph tend to cluster together and is based on triplets of nodes. A triplet is three nodes that are connected by either two (open triplet) or three (closed triplet) undirected ties and a triangle graph includes three closed triplets, one centered on each of the nodes. Having these, we can define the global clustering coefficient as:

$$Cl = \frac{3 \times \text{number of triangles}}{\text{number of all triplets}}, \tag{2.9}$$

$Cl = 1$ implies perfect transitivity, meaning that the network's components are all cliques and $Cl = 0$ implies no closed triads.

2.3 An Illustrative Example

Lastly, we will show a simple example in order to calculate some of the above metrics and get a deeper understanding of them. For this example, we will use an undirected network, which can be seen in Figure 2.3. Before starting to calculate different metrics, we must calculate the adjacency matrix. The adjacency matrix of the undirected network is the following:

$$A_{ij} = \begin{pmatrix} 0 & 1 & 0 & 1 & 1 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 \end{pmatrix}$$

Having the adjacency matrix, we can use it for calculating the network's metrics that we want. First, we will calculate the degree of each vertex using the formula in Eq 2.1. Thus, for the vertex 1 we will have:

$$k_1 = \sum_{j=1}^{6} A_{1j} = A_{11} + A_{12} + A_{13} + A_{14} + A_{15} + A_{16} = 3$$

Similarly, we can find the degree for the other vertices too (see Table 2.1). Next, we will calculate the eigenvector centrality for each vertex. Because, it requires to find the eigenvalues of the adjacency matrix, we will calculate it computationally. The results for each vertex can be seen in Table 2.1. Moreover, we will calculate Katz centrality using the Eq 2.4 and the results of each vertex are in Table 2.1.

Vertex	**Degree**	**Eigenvector**	**Katz**	**Closeness**	**Betweenness**
1	3	0.56	0.45	0.86	0.65
2	2	0.44	0.41	0.66	0.15
3	2	0.41	0.41	0.54	0.05
4	2	0.44	0.41	0.66	0.15
5	2	0.32	0.40	0.66	0.40
6	1	0.15	0.36	0.46	0

Table 2.1: Different network metrics for each vertex

The next metric that we will calculate is the Closeness centrality, using the Eq 2.7. So, for the vertex 3 we will have:

 https://doi.org/10.26219/heal.aueb.6624

$$C_3 = \frac{6}{\sum_j d_{3j}} = \frac{6}{d_{31}+d_{32}+d_{33}+d_{34}+d_{35}+d_{36}} = \frac{6}{11} = 0.54$$

Again, for the other vertices we can see the Table 2.1. The final metric that we will calculate is the Betweenness centrality. Using the Eq 2.8 we can find for vertex 4:

$$x_4 = \sum_{s \neq t \neq 4 \in N} \frac{n_{st}^{(4)}}{g_{st}} = \frac{4}{10} = 0.4$$

For all the other vertices we can see Table 2.1. Finally, we conclude that the vertex 1 has a significant role in the network, which can be seen from the high values that its metrics has compared to the ones of vertex 6 which doesn't play an important role in the network. The above example was just a simple illustration to see how we can calculate arithmetically some basic network metrics, for a network that has a lot of vertices the use of computer algorithms is necessary.

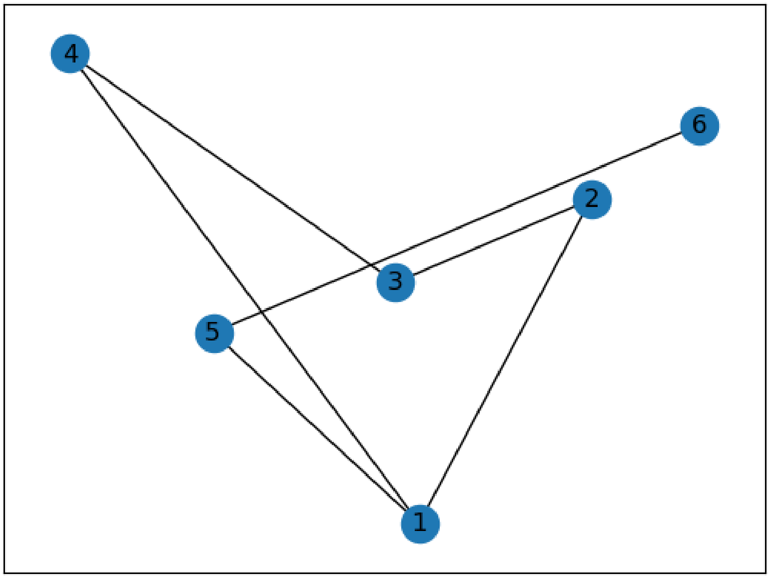

Figure 2.3: An undirected network with six vertices

https://doi.org/10.26219/heal.aueb.6624

Chapter 3

Literature Review

This chapter constitutes a comprehensive survey of relevant literature in the domain of financial networks. The first section delves into the concept of systemic risk, emphasizing the instrumental role played by network analysis in its understanding and mitigation. Subsequently, attention shifts to community detection within networks, exploring the identification of hidden substructures and applications in financial networks. Lastly, we present shareholders networks, aligning closely with the analytical focus of the thesis. By traversing these sections, readers gain invaluable insights from existing scholarship, laying the groundwork for the subsequent chapters and deepening our comprehension of financial networks.

3.1 Systemic Risk and Network Analysis

Over the past years there is a large contribution to the analysis and to the functional understanding of the structure of real-world networks. Boss (2003)[6] conducts an empirical analysis of the Austrian interbank market's structure and finds that the degree distribution of the interbank network shows two different power law exponents which are one-to-one related to two sub-network structures, differing in the degree of hierarchical organization. There are several studies that was done during these years but after the Banking crisis in 2008, network analysis started to attract a lot more attention. The concept "too-central-to-fail" was introduced by Battiston (2012) [7], where the authors proposed the DebtRank metric to identify nodes of systemic importance. The analysis that they did in the FED emergency loans program during the 2008-2010 financial crisis highlights the critical role of interconnected nodes in systemic risk. Additionally, Hideaki (2013) [8] delves into lending/borrowing relationships in the Japanese credit network, where they intro-

https://doi.org/10.26219/heal.aueb.6624

duce two different DebtRank metrics to assess node importance and vulnerability during crisis. Notably, the research reveals a non-linear connection between bank size and importance. Moreover, Thurner (2013) [9] addresses the lack of accurate risk assessment among banks within interbank networks. The authors propose introducing transparency by revealing each bank's DebtRank and incentivize borrowing from safer lenders. This self-organized critical approach redistributes risk and has the potential to significantly reduce systemic risk without compromising network efficiency. Furthermore, Bardoscia (2015) [10] establishes a dynamic "microscopic" theory of instability in financial networks. Building upon the DebtRank algorithm, the researchers generalize the model, providing economic intuition based on accounting principles. Their approach uncovers network effects amplifying exogenous shocks and emphasizes the importance of regulatory measures in addressing systemic risk within financial networks. Another interesting view is the one of Silva (2017) [11], where innovative network-based measurements are introduced to identify contagion roles in the Brazilian financial market. The impact susceptibility index assesses vulnerability to local or remote impacts and the impact diffusion influence index captures the potential influence of financial institutions in propagating impacts. Surprisingly, some non-large banking institutions consistently exhibit more influence in impact diffusion than large banks. Chabot (2019) [12] examines financial market stability for systemically important banks (SIBs). Proposes a fresh approach to comprehend banks' interactions in a complex environment, using network analysis on 262 European banks. The study discovers that a bank's network position and its local characteristics impact system stability more than its systemic importance. The research challenges the notion that SIBs with increased capital can solely ensure crisis containment and suggests integrating interconnectedness metrics into monitoring systems for regulators to identify potential disruptions early. A different approach of assessing systemic risk can be done using a multilayer network. Cuba (2021) [13], presents a comprehensive study of Peru's banking system, investigating interbank exposures through a multilayer network analysis. By considering various layers of financial interactions and employing metrics like DebtRank, the study assesses systemic importance while accounting for interconnectedness. Notable findings include a concentrated interbank network, a core-periphery structure in the multilayer analysis, and an inverse relationship between impact and vulnerability. Lastly, I.-E.Téllez-León (2021) [14], employs network theory, econometric models and machine learning to analyze the structural properties of the secured and unsecured interbank markets in Mexico. The study identifies key network measures influencing lender-borrower relationships, with centrality metrics like PageRank explaining secured market dynamics and systemic risk indicator DebtRank driving unsecured market trends. Findings highlight the sig-

https://doi.org/10.26219/heal.aueb.6624

nificance of a bank's position in the network, the systemic relevance of counterparts, and the density of connections in influencing funding prices.

3.2 Community Detection in Networks

Community structure detection in networks has been a topic of significant research interest due to its applications in various domains. Several methods have been developed to uncover and analyze community structures with different types of networks. Girvan (2001) [15] introduced a novel approach for identifying community structures by utilizing edge betweenness information. Unlike conventional methods that focus on identifying strongly interconnected core communities, Girvan's method capitalizes on detecting community peripheries. Through experiments on computer-generated and real-world networks, the approach demonstrated remarkably accuracy in identifying communities. Newman (2004) [16] reviews some traditional methods for community detection such as spectral bisection method, which is based on the eigenvector of the graph Laplacian, the Kernighan-Lin algorithm and hierarchical clustering. He mentions that these methods are unsuitable for real-world network data and then introduces some suitable methods based on iterative removal of between-community edges, including the betweenness-based method of Girvan and Newman, the Monte Carlo resampled variation proposed by Tyler, and the algorithm based on counts of short loops proposed by Radicchi. Finally, discuss two more algorithms for their computational efficiency, namely modularity maximization of Newman and resistor network algorithm of Wu and Huberman. Next using the findings of the previous authors, Leicht (2007) [17] proposes a method for finding communities in directed networks. They show that modularity can be generalized in a principled fashion to incorporate the information contained in edge directions, using an algorithm based on the eigenvectors of the corresponding modularity matrix. The method is an extension of the modularity optimization method for undirected networks. Their algorithm can detect also more than two communities using repeated bisection until the modularity of the network stops increasing. Furthermore, Blondel (2008) [18] propose a community detection method (Louvain method) based on modularity optimization, which is shown to outperform all other known methods in terms of computation time. Firstly, the method assigns a different community to each node of the network and then for each node i, evaluates the gain of modularity that would take place by removing the node i from its community and by placing it in the community j. The change happens when the gain is maximum. This process is applied repeatedly and sequentially for all nodes until no further improvement can be achieved. Another interesting algorithm came from Rosvall (2007) [19], where introduces a new

https://doi.org/10.26219/heal.aueb.6624

information theoretic approach that reveals community structure in weighted and directed networks. The method decomposes a network into modules by optimally compressing a description of information flows on the network. The method uses the map equation, which calculates the minimum description length of a random walk on the network for a two-level code that separates the important structures from the insignificant details based on the partition. Finally, two interesting applications in financial networks were done by Vitaly (2014) [20] and Huang (2021) [21]. Vitaly (2014) present an analysis of the global ownership network of transnational corporations using community detection methods such as the Louvain algorithm of Blondel (2008) and the Infomap method of Rosvall (2008). They conclude that the global corporate network is strongly clustered in communities, where geography is the major driver while sector is not so important. On the other hand, Huang (2021) uses the method of Blondel (2008) for community division in order to deal with large-scale networks of bank correlation and filters the divided communities into sub-networks. With that, he studies the differences in risks between different sub-networks and focuses on the relationship between the systemic risk of community sub-networks and community structural variables, such as community size.

3.3 Shareholders Networks

Shareholders networks play a critical role in modern financial markets, influencing information flow, control dynamics, and investment behaviors among market participants. Firstly, Souma (2005) [22] present the Japanese shareholding network existing at the end of March 2002. They consider a directed network by drawing edges from shareholders to stock corporations. Studying the outgoing degree distribution, they found that it follows the power law with an exponential cutoff. Also, from considering the correlations between the outgoing degree and the company's age, profit, and total assets, conclude that an important factor in the growth of the business network is not the company's age but its total assets. Next, M. D'Errico (2009) [23] studied the Shareholding Network embedded in the Italian Stock Market. They used different centrality measures in order to identify the central companies both in the role of transferring information flows and controlling companies. Specifically, suggest that betweenness centrality can identify companies which are central in the information flow and because the control of a company can be exercised in an indirect way, they propose the flow betweenness for detecting it. On the other hand, Li (2014) [24] construct a bipartite network of the energy listed companies and their shareholders and a derivative holding-based network based on the equivalence network theory. Also, propose the weighted and un-weighted Shareholding Similarity Coefficient, which

https://doi.org/10.26219/heal.aueb.6624

represent the similarity of the two sets of listed companies whose stocks are held by any two shareholders, regardless of the shareholding rate. They found that a relatively high shareholding similarity coefficient corresponds to relatively low numbers of holding and co-holding of the listed energy companies. A different approach was done by Yao (2019) [25], where they study the investment characteristics of different type of shareholders in Turkish and Dutch companies, using a network (for each country) of shareholders linked if they have invested in the same company. They calculate different network measures such as degree, assortativity, diversity of neighbors, betweenness and closeness centrality. Also, perform percolation analysis focusing on one type of shareholder and community detection. They found that the individual or family investors in Turkey is more peripheral than that of the Netherlands and the Corporate shareholder in Turkey form a central core for that network. Additionally, Long (2022) [26] propose a framework for quantifying relational risk based on publicly available risk events for small- and medium- sized enterprises (SMEs') credit risk evaluation. They quantify relational risk by weighting the impact of publicly available risk events of each firm in shareholding and governance networks. In order to do this, they use a smoothed version of the wvRN algorithm to generate a score by weighting the time dependent risk scores of firms based on the network structure and combine it with basic features to evaluate the credit risk of SMEs. They found that the additional information of relational score in the models, improves both discrimination and granting performances of credit risk evaluation of SMEs. Finally, Liu (2023) [27] present a shareholder-stock bipartite network and analyze the topology of the network to explore the structural characteristics of overlapping portfolios among different shareholders, as well as similar shareholder allocation structures between different stocks. For the analysis, they use the method of a statistically validated network model to establish the projected networks, then they calculate basic network measures such as degree, clustering coefficient and PageRank and finally perform community detection. Lastly, they conclude that most shareholders have different shares and investment portfolios.

https://doi.org/10.26219/heal.aueb.6624

Chapter 4

Understanding the Network: Data and Analysis

In this chapter our aim is to understand the relationships between shareholders and the companies they invest in. In order to achieve this we will do a comprehensive exploration of the bipartite network, which will help us unveil the complex connections within this network. After that, we will delve into the concept of projected networks and split the bipartite network into two separate unipartite networks representing shareholders and companies. This separation will enable us to apply traditional network analysis techniques in both projected networks, providing deeper insights into the financial relationships. However, before starting our analysis we will make a small presentation of the datasets that we will use.

4.1 Data Overview

In this section, we present a comprehensive examination of the datasets employed in our analysis. Our data encompasses two primary synthetic datasets: one includes the shareholders and the companies that they have invest, while the other one consists information about if a company had a default event or not in different time periods. It is important to note that these synthetic datasets were carefully prepared by Ernst&Young Credit Risk Analysts, who incorporated expert judgment, insights, and observations derived from real-life data.

More specifically, the first dataset includes 1277 rows and two columns namely shareholder_id and company_id. Each row of the dataset represents the shareholder and the company that he has invest. In total we have 323 different shareholders and 830

https://doi.org/10.26219/heal.aueb.6624

different companies. In the context of network analysis, this dataset can be represented as a bipartite network consisted two sets of nodes, one with the shareholders and the other with the companies and an edge between a shareholder and a company that he has invest.

On the other hand, the second dataset includes 19921 rows and three columns namely company_id, target and reference date. The company_id column signifies the distinctive identifiers assigned to companies. The target column, a binary variable, assumes a value of 1 to denote a default event or 0 for a non-default event. The reference date column runs on a quarterly basis from 2017 to 2022 and specifies the precise date attributed to the companies' financial assessments. This dataset is used for credit risk assessment using predictive network metrics.

All in all, the first dataset serves as the foundation for our network construction, forming the basis upon which we create a comprehensive representation of shareholder-company interactions. This network enables us to explore the complex relationships within our financial ecosystem. In contrast, the second dataset is used for assessing the predictive power of newly introduced network metrics. By applying these metrics to the second dataset, we endeavor to unveil their potential in forecasting binary outcomes, specifically default events, contributing valuable insights to our credit risk assessment framework.

4.2 Bipartite Network Analysis

In this section we will investigate the bipartite network, building upon our understanding of what a bipartite graph is as previously outlined in chapter 2. The bipartite network under study represents the relationships between shareholders and the companies they invest in. Our network consists of 1153 nodes and 1276 links and this makes the visualization of the network uninformative. If we want to get an idea of how the visualization of the network will be, we can look the Figure 2.1. Now, we can turn our attention to specific network measures that will help us extract valuable information from the network.

4.2.1 Degree Centrality

Degree centrality is a fundamental network measure that quantifies the importance of a node within the network by counting the number of edges connected to that node. In the context of bipartite network we introduce degree centrality for both

https://doi.org/10.26219/heal.aueb.6624

shareholders and companies. Bipartite graphs, as we have mentioned, have two sets of nodes S and C with the shareholders and the companies respectively and one set E with edges that only connect nodes from opposite sets. Furthermore, the maximum possible degree of a node is the number of nodes in the opposite node set. So, the degree centrality for a node v in the bipartite sets S with n nodes and C with m nodes is [28]:

$$\begin{aligned} d_v &= \frac{deg(v)}{m}, for v \in S \\ d_v &= \frac{deg(v)}{n}, for v \in C \end{aligned} \tag{4.1}$$

For example, in our case the degree centrality for shareholder $s1$ and for the company $c1$ will be:

$$\begin{aligned} d_(s1) &= \frac{deg(s1)}{830} \\ d_(c1) &= \frac{deg(c1)}{323} \end{aligned}$$

where 830 is the number of companies in the node set C and 323 the number of shareholders in the node set S.

Now, using the Eq 4.1, we can calculate the degree centrality for each node and make a histogram in order to understand the distributions, but because we have a lot of nodes the degree centrality will have very small values and will make it harder to interpreted. So, it is better to take the histograms of the degrees. Looking at

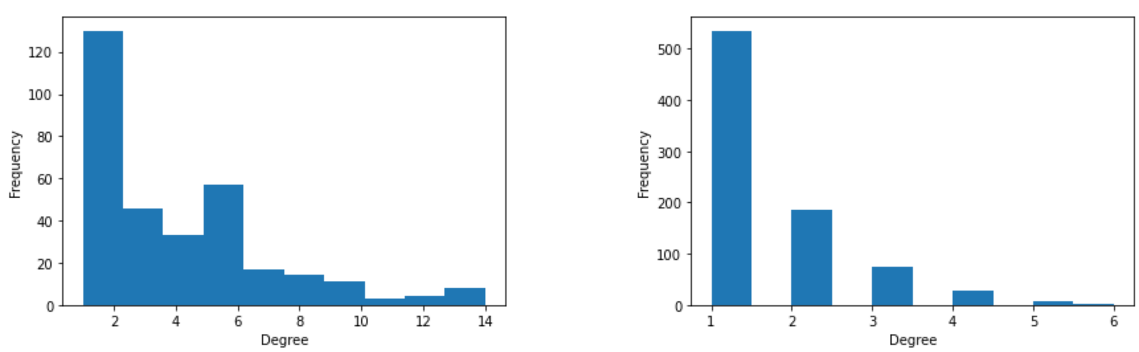

Figure 4.1: Histograms of the degrees for the shareholders and the companies respectively.

 https://doi.org/10.26219/heal.aueb.6624

the Figure 4.1, we can see that both histograms are right skewed. Regarding the histogram with the shareholders, we can conclude that there are 3 different groups of shareholders. The first one consists shareholders that have invest in one company, the second one those that have invest between 2 and 5 companies and the last one which is also the smallest one includes shareholders that have invest in more than 6 companies. This suggests that most of the shareholders have a more conservative investment strategy compared to the other that maintain a broader and more diversified portfolios. On the other hand, from the histogram with companies, we can see that most of them have only one shareholder and only one company has 6 shareholders.

4.2.2 Betweenness Centrality

The next measure that we will introduce is the betweenness centrality for bipartite graphs. As we mentioned previously in chapter 2, betweenness centrality measures the number of times a node lies on the shortest path between other nodes. More specifically, betweenness centrality of a node v is the percentage of all the shortest paths of any two nodes (apart from v), which pass through v. In the case of bipartite graphs the formula is slightly different from the one in the unipartite case, because bipartite graphs have two distinct sets of nodes and we need to consider paths that connect nodes from one set to nodes in the other set. So, the formula for the betweenness centrality of a node v in a bipartite graph is defined as:

$$C_B(v) = \sum_{s \in S, c \in C} \frac{\sigma_{sc}(v)}{\sigma_{sc}}, \tag{4.2}$$

where S and C are the two disjoint sets of nodes, s and c are nodes from the sets S and C respectively, σ_{sc} is the total number of shortest paths from a node in set S to a node in set C and $\sigma_{sc}(v)$ is the number of those shortest paths that pass through node v.

Now, that we have the formula, the next step is to use it for calculate the betweenness centrality for every node in our graph and make two histograms for the shareholders and the companies to get a better view of their distribution of betweenness centralities. Before, we present the two histograms, we have to mention that the values of betweenness are normalized by the maximum possible value which for bipartite graphs is limited by the relative size of the two node sets [28].

https://doi.org/10.26219/heal.aueb.6624

Let n be the number of nodes in the node set S and m be the number of nodes in the node set C, then nodes in S are normalized by dividing by:

$$\frac{1}{2}[m^2(s+1)^2 + m(s+1)(2t-s-1) - t(2s-t+3)],$$

where

$$s = (n-1) \div m, t = (n-1) \mod m,$$

and nodes in C are normalized by dividing by:

$$\frac{1}{2}[n^2(p+1)^2 + n(p+1)(2r-p-1) - r(2p-r+3)],$$

where

$$p = (m-1) \div n, r = (m-1) \mod n$$

Looking at the Figure 4.2, we can get some insights into the roles and interactions of both shareholders and companies. As we can see in the left histogram, the majority of the shareholders have 0 centrality, which indicates their limited involvement in bridging connections between shareholders and companies. Moreover, we can see also a few shareholders that have values between 0.11 and 0.20, which suggests that they have more substantial roles in facilitating connections between different sets of companies. On the other hand, in the right histogram with companies we can see a similar pattern. Most of the companies have very low centrality values, suggesting that they do not significantly mediate connections with shareholders. However, a minority of companies display moderate centrality, indicating their importance in bridging different shareholder groups.

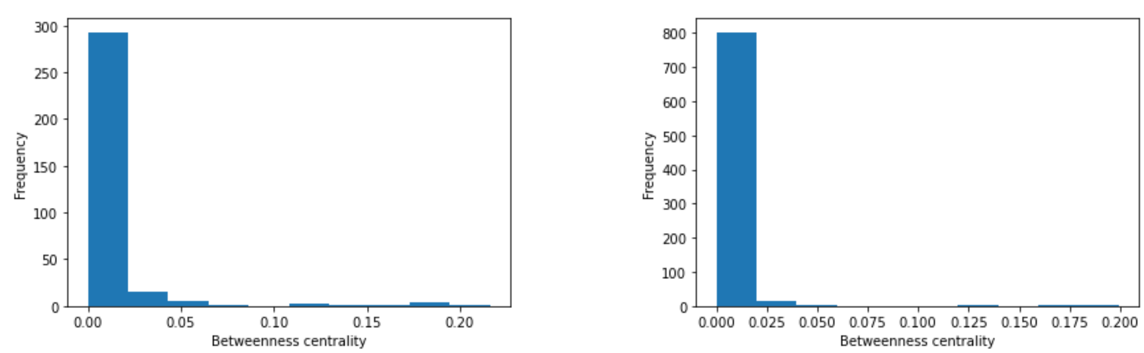

Figure 4.2: Histograms of the betweenness centralities for the shareholders and the companies respectively.

4.3 Unipartite Network Analysis

In this section we will try to analyze the network using the measures that we introduced in chapter 2, which is a contradiction because those measures are for unipartite networks. For that reason, we will begin with an introduction to projected networks, a powerful concept that will help us transform the bipartite network into two different unipartite networks. Following this, we will delve into the projected network analysis where we will analyze the two unipartite networks separately using traditional network analysis techniques. Through this analysis, we aim to unveil hidden structures, identifying influential nodes and illuminate the underlying dynamics that define the relationships between shareholders and companies in our network.

4.3.1 Projected Networks

The method that we will employed is called shared neighbor method. This method will help us create two unipartite networks. The first one will be the Shareholder-Shareholder (S-S) network which highlights the relationships among shareholders, based on their shared interactions with companies. The second one will be the Company-Company (C-C) network, which focuses on the companies based on their common shareholders.

In order to understand exactly what the method does, we will do a simple example. We will use a bipartite network with 3 shareholders (blue color) and 4 companies (green color), as we can see in Figure 4.3. Using the projected method, we will take two networks, one for the projection of the vertex set S with respect to the vertex set C and the other one for projecting the vertex set C with respect to the vertex set S. Regarding the first projection, we can see that both shareholders 's1' and 's2 have invested in the company 'c1', which means that they have at least one shared neighbor. This creates a link between them in the projected network. Also, we can see that 's1' and 's3 do not have invest in the same companies, which results to not have any links between them in the projected network. On the other hand, companies 'c3' and 'c4 have a similar shareholder 's3', which makes them have a link in the projected network.

https://doi.org/10.26219/heal.aueb.6624

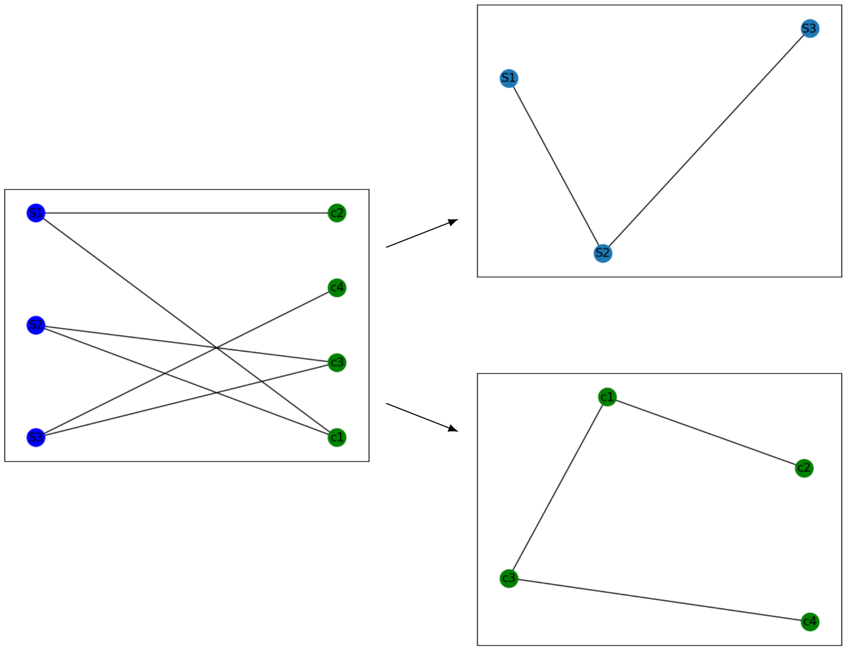

Figure 4.3: Bipartite network on the left and its two projections on the right.

4.3.2 Shareholder-Shareholder Network Analysis

Having in mind the example that we did, we will use the same philosophy for constructing the Shareholder-Shareholder network. The network consists of 323 nodes as shareholders and an edge between two shareholders if they have invested in at least one common company. The network is depicted in Figure 4.4, where we can see that we have two groups. The first group consists shareholders that they have 0-2 edges and the second one consists shareholders with much more edges. Now, that we have the unipartite network, we will calculate different network measures like the ones that we introduced in chapter 2 and try to find more about the relationships among shareholders.

https://doi.org/10.26219/heal.aueb.6624

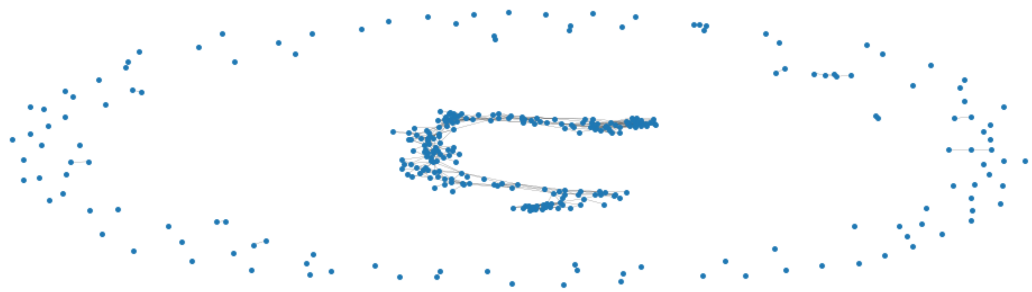

Figure 4.4: Shareholder-Shareholder network

The first measure that we will study is the degree centrality (i.e. the degree divided by the number of shareholders). In this analysis, degree centrality means that the higher the centrality of a node is, the more edges are connected to the particular node and thus the more neighbor nodes (common investments) this node has.

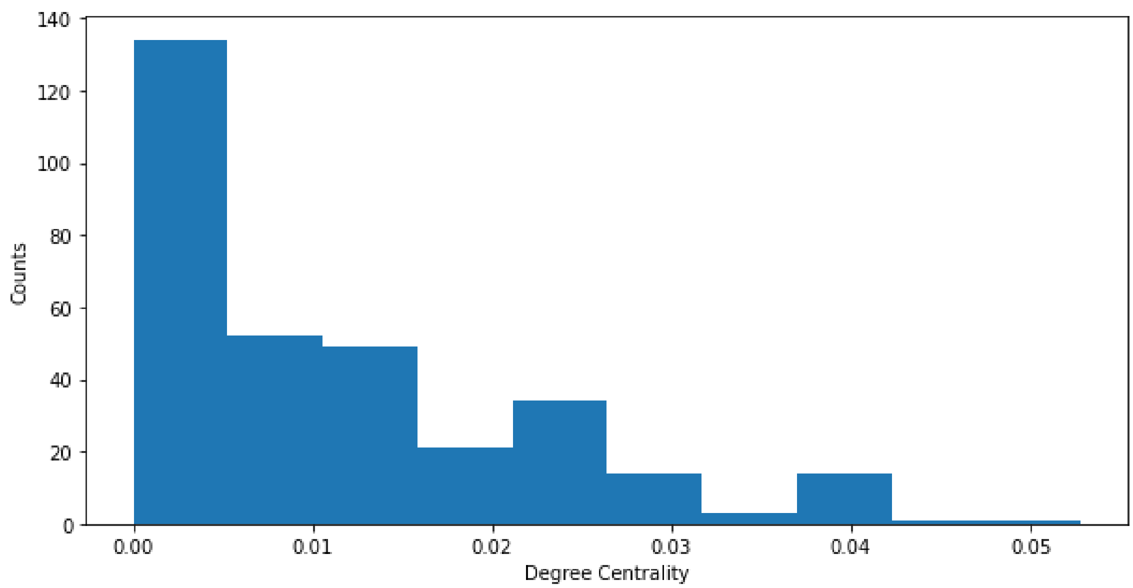

Figure 4.5: Degree centrality histogram of the S-S network.

It is visible from Figure 4.5, that the vast majority of the shareholders have degree centralities of less than 0.01 and a few reach the value of 0.05. We can verify this, looking at the Figure 4.4, where a lot of shareholders have 0-2 degree and some in the center have higher values of degree, which results to the distribution of degree centrality to be right-skewed. Also, calculating the average degree, we found that is

https://doi.org/10.26219/heal.aueb.6624

3.5 suggesting that, on average, each shareholder has invested in the same companies to approximately 3.5 other shareholders, which implies a moderate level of collaboration or investment overlap among shareholders.

Next, we will study the betweenness centrality which measures the number of times a node lies on the shortest path between other nodes, meaning it acts as a bridge. Specifically, in the S-S network this measure is associated with the shareholder's ability to influence others. For example high betweenness centrality can indicate that a shareholder holds a central position in the network and maybe sought after for partnerships or investment opportunities due to their ability to connect others. Moreover, as we can see from Figure 4.6, the majority of the shareholders centrality values are below 0.03, which makes sense as the network is very sparse and most of the nodes do not act as bridges. However, we can see that we have a very few shareholders with centrality values up to 0.14. These individuals play crucial roles in information flow and facilitating interactions.

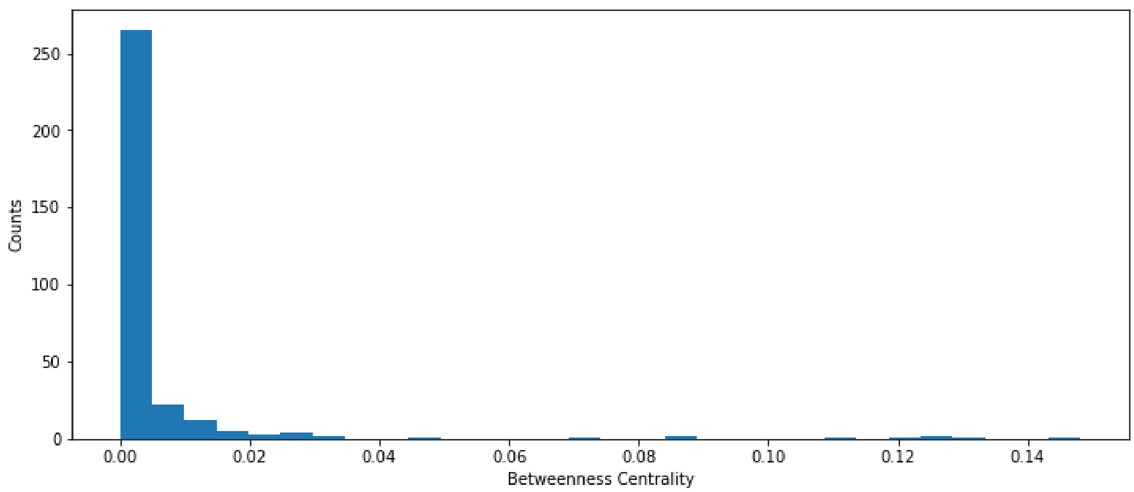

Figure 4.6: Betweenness centrality histogram of the S-S network.

Another important measure that we will calculate is the closeness centrality which measures the proximity of a node to all other nodes in the network. In other words, the higher the closeness centrality of a node is, the closer it is located to the center of the network. It is very important for the monitoring of the spread of information. If a shareholder with the highest closeness centrality measure started spreading some information, the whole network would get informed the quickest possible. As previously, we will look at the histogram of the closeness centrality in order to get a better understanding of its distribution. Looking at Figure 4.7, the presence of two groups

https://doi.org/10.26219/heal.aueb.6624

is notable. The first group, characterized by values ranging from 0 to 0.01, represents shareholders who are isolated from others. In contrast, the second group, with values from 0.07 to 0.14, consists of shareholders who are centrally located, with shorter paths to reach other shareholders. These individuals play pivotal roles in spreading information and have a higher potential for influence within the network.

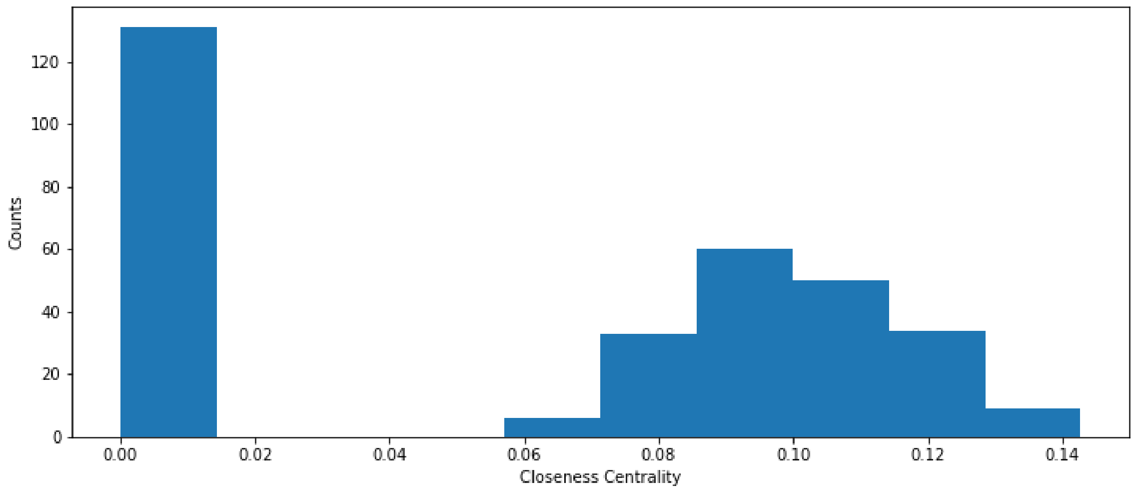

Figure 4.7: Closeness centrality histogram of the S-S network.

The last centrality measure that we will analyze is the eigenvector centrality, which shows how connected a node is to other important nodes in the network. It measures a node's influence based on how well it is connected inside the network and how many links its connections have and so on. Looking at the Figure 4.8 we can see that most of shareholders have eigenvector centralities 0.

https://doi.org/10.26219/heal.aueb.6624

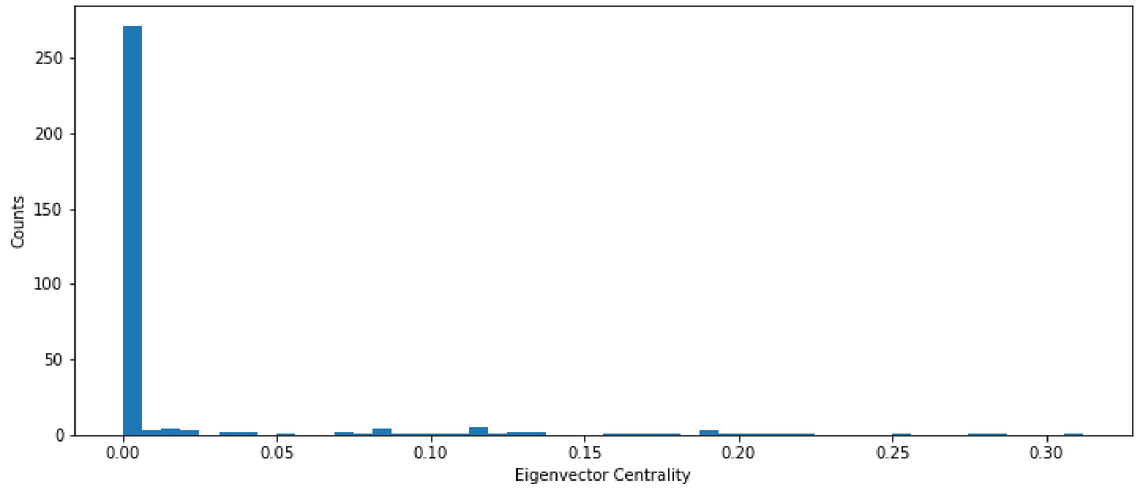

Figure 4.8: Eigenvector centrality histogram of the S-S network.

However, we can also see different values of eigenvector centralities as there are tiny bins all over the x axis. These shareholders that have values close to 0.30 can be considered the most important in this network in terms of overall influence to the whole network. They have the ability to shape network's dynamics and potentially influence the decisions and actions of other shareholders.

The last measure that we will see for the S-S network will be the clustering coefficient, which is not a centrality measure. Clustering coefficient measures the extent to which shareholders tend to form clusters or tightly-knit groups within the network. Specifically, it quantifies the likelihood that a shareholder's neighbors are also connected to each other. The interpretation of the clustering coefficient can be valuable for understanding the network's cohesion and identifying potential communities or cliques of shareholders.

https://doi.org/10.26219/heal.aueb.6624

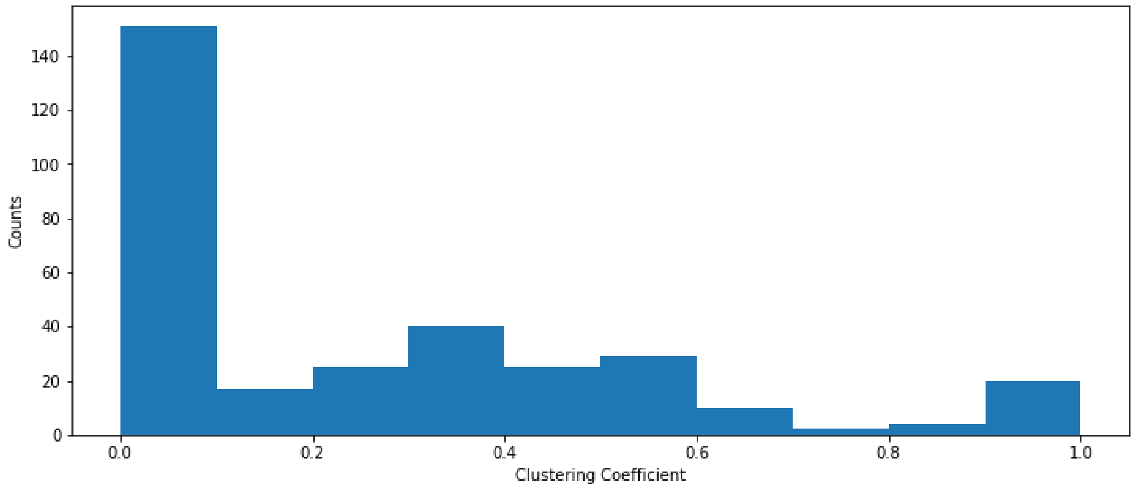

Figure 4.9: Clustering coefficient histogram of the S-S network.

From Figure 4.9, we can see that we have numerous values close to 0, which indicates that a lot of shareholders have limited interactions with others. This makes sense because as we saw previously the majority of shareholders have 0 degree. The presence of values in the range of 0.2 to 0.6 suggests that there are clusters or subgroups in the network who interact more closely with each other. These subgroups likely represent communities of shareholders with shared interests. Lastly, there are a few values equal to 1. This indicates that a shareholder and his neighbors are fully interconnected, forming a complete clique which might represent a specific subset of shareholders who are engage in extensive collaboration.

The final thing that we will do, is to find the top five shareholders based on different centrality measures and see if they have similarities or not. Looking at Table 4.1, we can see the shareholder id and the value of each centrality measure for each one of them. The maximum degree centrality is 0.05, which means that the shareholder 's178' is connected to approximately 5% of all other shareholders in the network and the remain shareholders to 4%. Besides having the highest degree, we have different shareholders with the highest betweenness centrality. This happens, because betweenness centrality emphasizes to the control of information flow between others, so it is possible for shareholders to have many connections (high degree) but not necessarily serve as bridges. On the other hand, we can see a lot of similarities between the shareholders with the highest betweenness and closeness centrality. This suggests, that the shareholders that have the ability to influence others are also positioned close to the center of the network. Furthermore, looking at the last column of the Table 4.1, we can see that only the shareholders 's178' and 's177' have be seen before and all the others are identified for the first time, meaning that they have neither the highest degree, betweenness or closeness centralities in the graph. That leads to the conclusion that those shareholders are very likely to be connected with the 's178' and as a result have very high eigenvector centralities. Indeed, the shareholders 's177','s193' and 's199' are connected with 's178'.

Degree		**Betweenness**		**Closeness**		**Eigenvector**	
Shareholder Id	Value	Shareholder Id	Value	Shareholder Id	Value	Shareholder Id	Value
s178	0.05	s79	0.15	s97	0.14	s178	0.31
s177	0.04	s97	0.13	s117	0.14	s197	0.28
s176	0.04	s58	0.13	s79	0.14	s177	0.27
s139	0.04	s117	0.12	s138	0.13	s193	0.25
s77	0.04	s138	0.12	s112	0.13	s199	0.22

Table 4.1: Top five shareholders for each network centrality measure.

4.3.3 Company-Company Network Analysis

Following similar steps as before, we will do now the analysis for the projected network of companies. In order to find the projection of the vertex set C (companies) with respect to the vertex set S (shareholders), we will use the method of shared neighbor. This means, that now each company becomes a node and an edge is drawn between two companies if they share at least one shareholder. From that, we build a unipartite network with 830 nodes and 3249 edges. A visualization of the network can be seen in Figure 4.10. As previously the graph is not very clear due to the high number of nodes, but we can detect again two groups of companies, where the first one consists companies that are isolated from others and the second one companies that are more connected.

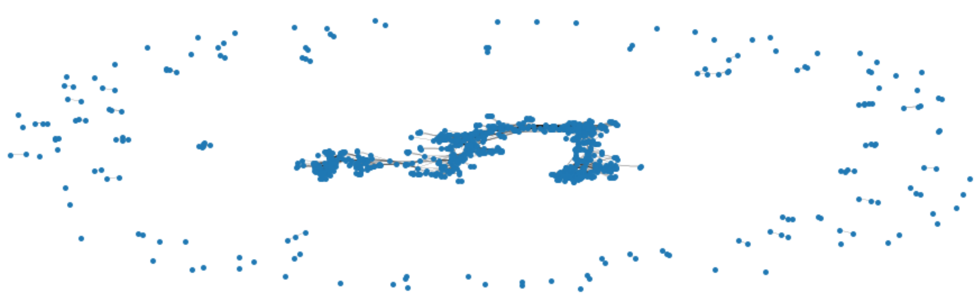

Figure 4.10: Company-Company network.

https://doi.org/10.26219/heal.aueb.6624

The first measure that we will look is the degree centrality. After calculating the degree centrality for each company we make a histogram in order to see the distribution. The histogram in Figure 4.11 reveals that the distribution is right-skewed. In general, the majority of companies have values 0 indicating that they don't have common shareholders with other companies and a few with values 0.05 suggesting that are more connected with others and probably serving as pivotal players in the network. Also, the average degree of the network is 8, which demonstrates a notable level of interconnectedness among companies.

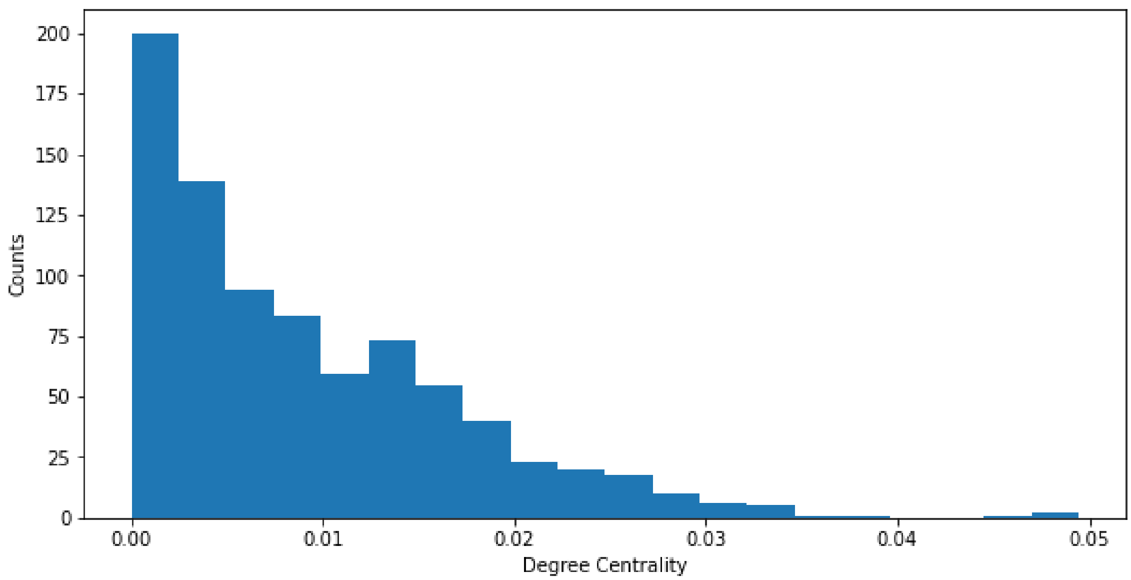

Figure 4.11: Degree centrality histogram of the C-C network.

Next will be betweenness centrality, a key metric, which quantifies the extent to which individual companies act as intermediaries or bridges facilitating connection between other companies. Looking at Figure 4.12, most values clustered around 0 and only a few reach as high as 0.25 suggesting that the majority of companies play limited role as intermediaries in connecting others. However, the presence of a few companies with notably higher betweenness centrality, have the potential to influence the network's stability. A financial shock or default in one of these companies could have huge consequences, propagating risk to other connected companies.

https://doi.org/10.26219/heal.aueb.6624

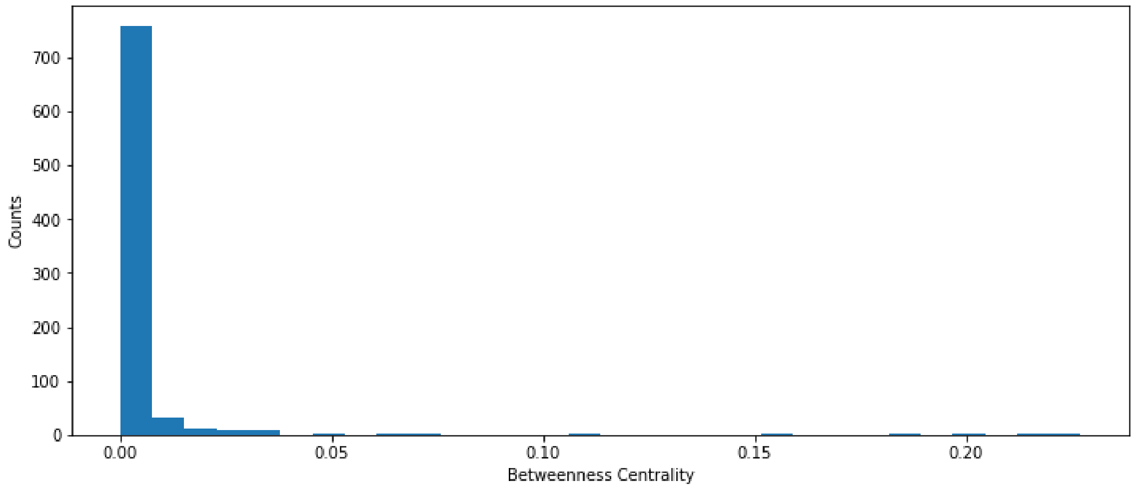

Figure 4.12: Betweenness centrality histogram of the C-C network.

Another important measure is the closeness centrality, which measures how quickly a company can reach other companies in the network. Companies with high closeness centrality have shorter average distances to all other companies, making them well positioned for rapid communication and interactions within the network. In the histogram of Figure 4.13, most of the companies have centrality 0 with a smaller group from 0.08 to 0.18. If one of the companies with high closeness centrality face a financial distress, the risk would be spread rapidly to other connected companies. Therefore, it is crucial to identify and closely monitor these highly central companies.

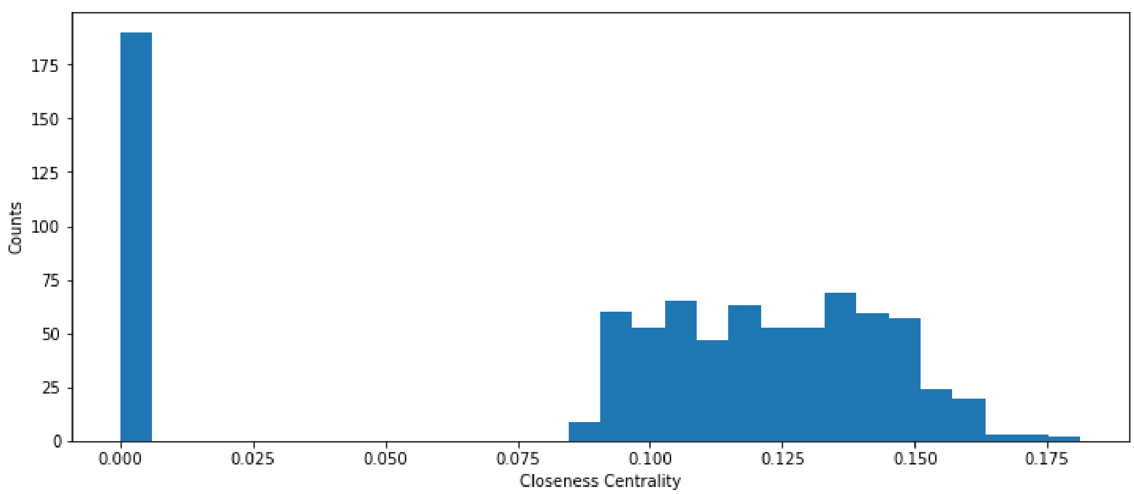

Figure 4.13: Closeness centrality histogram of the C-C network.

https://doi.org/10.26219/heal.aueb.6624

Furthermore, we will calculate the eigenvector centrality for each node, which assesses the influence and importance of each company based not only on its direct connections but also on the quality and significance of its connections. Companies with high eigenvector centrality are not only well-connected but are connected to other well-connected companies, signifying their influential positions in the network. Looking at the Figure 4.14, we can see that the majority of companies have value 0 and a few others have values up to 0.25. These companies can be considered important in this network in terms of overall influence to the whole network.

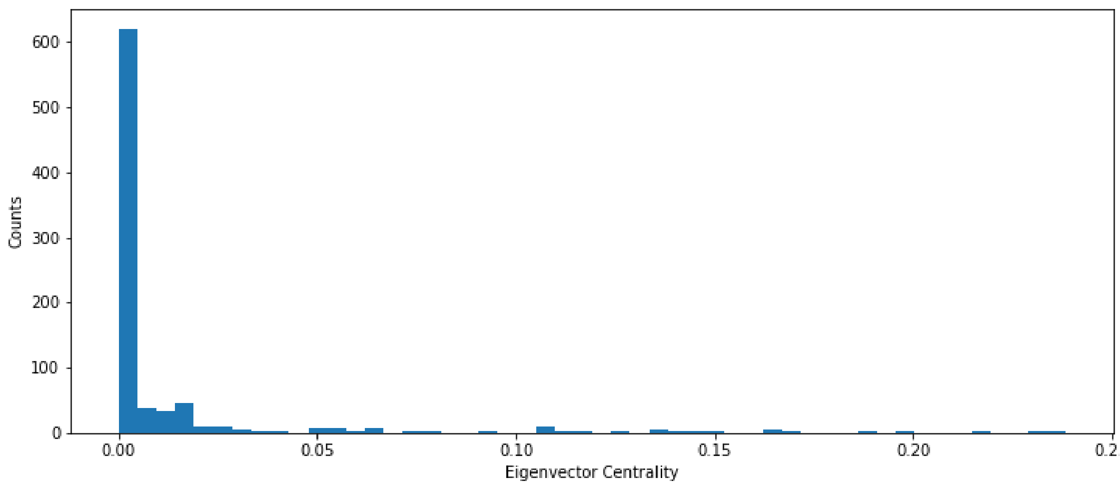

Figure 4.14: Eigenvector centrality histogram of the C-C network.

Additionally, the study of clustering coefficient will provide a better understanding of network's structure. The histogram in Figure 4.15 reveals a diverse distribution of clustering coefficients, with a substantial number of companies having a clustering coefficient equal to 1. This suggests the presence of tightly interconnected clusters or subgroups of companies within the network, which can be seen as risk concentrations. If a financial issue were to affect one company within such a cluster, it could potentially propagate across others. Also, the presence of companies with clustering coefficients between 0.2 and 0.8 indicates varying degrees of local cohesion and collaboration. Companies with values 0 signify limited local clustering or partnerships within their immediate network neighborhoods. This spectrum of clustering coefficients underlines the network's complexity, with both high clustered and less cohesive segments.

https://doi.org/10.26219/heal.aueb.6624

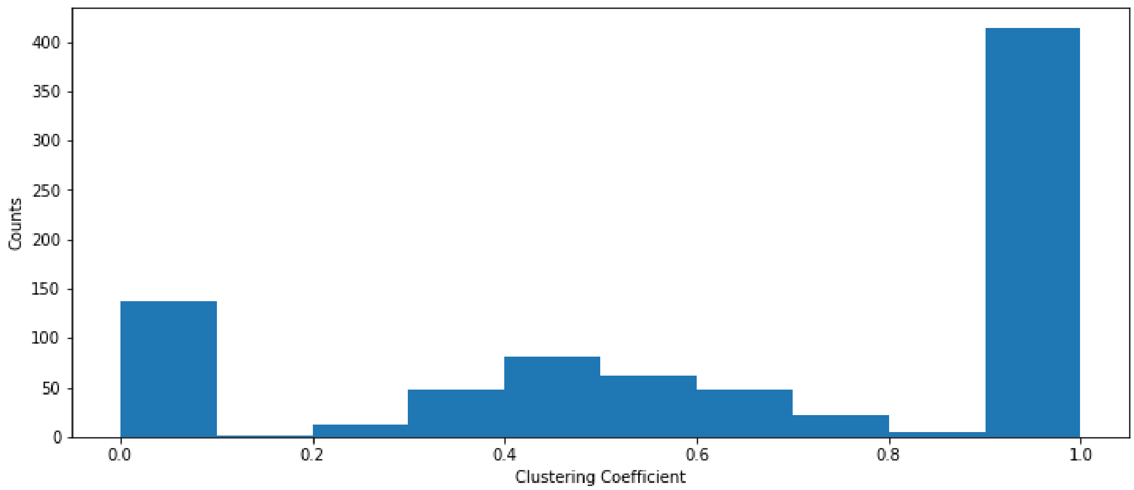

Figure 4.15: Clustering coefficient histogram of the C-C network.

The last thing that we will do in this analysis is to find the lengths of the shortest paths that connect any node to another node in the network. This requires computing the shortest path between every pair of nodes in the network, which will be done with a built in function in the NetworkX package. After calculating the lengths of the shortest paths, we make a histogram of the shortest path lengths which can be seen in Figure 4.16. Also, we have to mention that we did not account for the nodes that have 0 degree, resulting in no paths to other nodes. The distribution of shortest path lengths in the C-C network, ranging from 1 to 15. The histogram reveals the varying distances between companies within the network, with the majority of paths falling within 4 and 6. Shortest paths of length 1 indicate direct connections, reflecting shared shareholders or established partnerships. As the path lengths increase beyond the mean, we encounter more complex routes with a maximum value of 15. From a risk perspective, understanding the distribution of shortest path lengths is crucial for assessing the potential transmission of financial risks within the network and we will see in the next chapter how we can use this for creating a new metric that will help us predicting default events. Shortest paths of length 1 suggest that certain companies are tightly interconnected, which could amplify the impact of financial stress or defaults. Conversely, longer paths may provide a degree of resilience, as they require multiple intermediary connections to propagate risks.

https://doi.org/10.26219/heal.aueb.6624

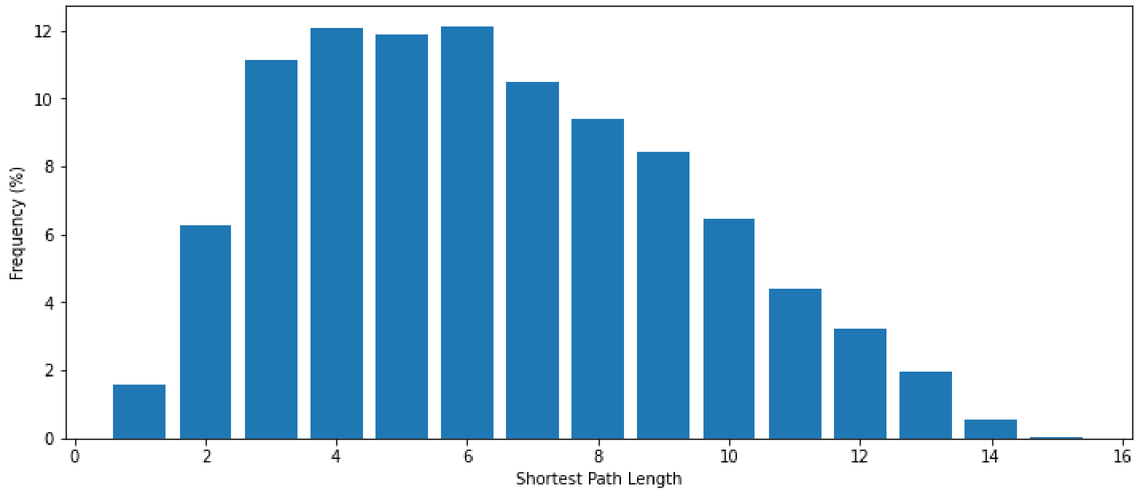

Figure 4.16: Distribution of shortest path length in the C-C network.

Finally, we will find the top five companies based on different centrality measures and see if we have similarities or not as we did previously for the S-S network. Looking at the Table 4.2, we can see a similar pattern as the one we had for the shareholders. The company 'c412' has the maximum degree centrality equal to 0.05. Regarding the ones with the highest betweenness and closeness centrality, we have the same companies except one. These companies are crucial to monitor because are central in the network and can influence the network's stability. So, in the case that one of them faces a financial shock, the event can spread rapidly and affect all the other companies that are connected with. Lastly, the companies with the highest eigenvector centrality are not the same with the others except 'c675' which is also fifth in the degree centrality column. A further investigation, saws that these companies are connected with the 'c604' resulting in an increase in their score.

Degree		**Betweenness**		**Closeness**		**Eigenvector**	
Company Id	Value	Company Id	Value	Company Id	Value	Company Id	Value
c412	0.05	c581	0.23	c581	0.18	c604	0.24
c449	0.05	c489	0.21	c489	0.18	c618	0.23
c528	0.04	c368	0.20	c674	0.17	c574	0.22
c1	0.04	c674	0.19	c368	0.17	c675	0.20
c675	0.03	c269	0.18	c512	0.17	c658	0.19

Table 4.2: Top five companies for each network centrality measure.

https://doi.org/10.26219/heal.aueb.6624

Chapter 5

Contributions and Results

In this chapter, we introduce two innovative network metrics namely "company score" and "distance from default", aimed at enhancing the credit risk assessment within financial networks. Each metric is defined in detail to provide a comprehensive understanding of its significance and application. To make it easier, we present illustrative examples that shed light on the metrics' functionality and relevance in real-world financial scenarios. Subsequently, we proceeded to construct three logistic regression models, each utilizing distinct predictors: one employing traditional network variables, another incorporating the company score, and a third integrating the distance from default metric, with the objective to assess the predictive potential of these metrics relative to the traditional network variables model. To compare these, we employed various performance measures including recall, accuracy, F1-score, and the Gini coefficient, widely recognized in credit risk assessment.

5.1 Novel Network Metrics

In this section we introduce a set of innovative network metrics namely "company score" and "distance from default", tailored to illuminate the complex dynamics within the shareholder-company network. Company score metric offers fresh insights into how a company can be affected by its shareholders investment behaviour and distance from default metric suggests that the distance from a company with a default event can influence the state of a company. Both metrics offer a valuable perspective on credit risk assessment that goes beyond traditional financial analysis.

https://doi.org/10.26219/heal.aueb.6624

5.1.1 Company Score

The company score metric is a powerful tool for evaluating the credit risk of companies within a shareholder-company network. It offers a comprehensive assessment by considering the influence of their shareholders and their investments in companies that have experienced default events.

In order to define the company score, we have to introduce first the shareholder score. The shareholder score is like a weighted average which assigns a value from 0 to 1 in each shareholder based on his investments. This score is given by the following formula:

$$Shareholder\_score_i = \frac{1}{Z_i} \sum_{j \in N(i)} w_{ij} p(T_j), \tag{5.1}$$

where Z_i represent the number of the total companies that shareholder i has invest, $N(i)$ are the neighbors of shareholder i in the network, w_{ij} denote the weight associated with the link between shareholder i and company j, which in our case is set to 1, $T_j = \{\text{Default Event, No Default Event}\}$ and the probability of a company j, $p(T_j)$, is set to 0 or 1 depending on whether the company j has a default event.

Now that we have the score for each shareholder we can use it to define the company score as following:

$$Company\_score_i = \frac{1}{S_i} \sum_{j \in N(i)} Shareholder\_score_j, \tag{5.2}$$

where S_i is the total number of shareholders that the i company has, $N(i)$ are the neighbors of company i in the network and $Shareholder\_score_j$ is the score for each shareholder which is calculated by the Eq 5.1.

For understanding the company score metric better, we will make a simple example in order to see how we can calculate it. Let's consider the data in Table 5.1, which is in a similar format as the dataset that we used for our analysis in chapter 4.
Each row in the Table 5.1 represents the shareholder, the company that he has invest and if the company had a default event. Using the Eq 5.1 and 5.2 we can calculate the company score of C_1 as follows:

$$Shareholder\_score_1 = \frac{1}{3}\Big(1p(T_1) + 1p(T_2) + 1p(T_3)\Big) = \frac{1}{3}(1 + 0 + 1) = \frac{2}{3}$$

https://doi.org/10.26219/heal.aueb.6624

Shareholder Id	**Company Id**	**Target**
S_1	C_1	1
S_1	C_2	0
S_1	C_3	1
S_2	C_1	1
S_2	C_2	0

Table 5.1: Shareholder-Company relationships

$$Shareholder\_score_2 = \frac{1}{2}\Big(1p(T_1) + 1p(T_2)\Big) = \frac{1}{2}(1+0) = \frac{1}{2}$$
$$Company\_score_1 = \frac{1}{2}\Big(\frac{2}{3} + \frac{1}{2}\Big) = \frac{7}{12}$$

In summary, the company score metric provides a comprehensive and network-informed perspective on credit risk. By providing a single numerical score, the metric offers a clear and concise representation of a company's credit risk. A higher company score suggests a potentially higher credit risk. Companies with elevated scores may warrant closer monitoring and proactive risk mitigation measures.

5.1.2 Distance from Default

The "Distance from Default" metric emerges as a valuable tool designed to navigate the complex landscape of financial networks. It addresses the fundamental question: how close or distant is a specific company from others in the network that have experienced default events? By quantifying this proximity, the metric offers unique insights into credit risk assessment within the context of the shareholder-company network.

The rationale behind this metric lies in recognizing that a company's financial health is not determined in isolation. Instead, it is intricately linked to the health of its network peers. Companies are interdependent, and bad financial events in one part of the network can spread through connected entities. Therefore, calculating a company's distance from those that have faced defaults provides a complete view of credit risk, incorporating both its financial condition and its network context.

Next, we will present the algorithm for calculating the distance from default metric. As we can see from Algorithm 1, in order to calculate the distance from default, we need the company projected network and a dataset that contains the companies id

https://doi.org/10.26219/heal.aueb.6624

and if they had a default event (target).

Algorithm 1: Distance from default

Input: Graph representing the company projection network.
Output: short_paths (a list containing shortest path lengths for each node).
Data: Dataframe with 'company_id' and 'Target' columns.

```
all_pairs=nx.all_pairs_shortest_path_length(network)
short_paths=[ ]
for each source_node in df['company_id']: do
    for each target_node in df where 'Target'=1: do
        if source_node!=target_node: then
            if target_node is present in all_pairs for source_node then
                short_path_length=all_pairs[source_node][target_node]
            else
                short_path_length=20 (or any suitable default value)
            short_paths=(source_node,short_path_length)
```

First, we use the NetworkX package in Python for calculating the shortest path lengths between all nodes in the graph. In order to do this we use the function all_pairs_shortest_path_length, where as an input it takes the network as a graph object. This function uses the Breadth-first search (BFS) algorithm in order to find the shortest paths between all pairs of nodes in the graph. In Algorithm 2, we can see how the BFS algorithm is employed to find the shortest paths from a selected starting node to all other nodes in a graph. The algorithm begins by initializing a queue, a set of visited nodes, and dictionaries to store distances and parent-child relationships (the tree structures). The starting node is enqueued and marked as visited, with its distance set to 0 and the parent set to None. The algorithm then iteratively explores nodes in a breadth-first manner, dequeuing each node and examining its neighbors. If a neighbor has not been visited, it is enqueued, and its distance and parent information are updated. This process continues until all reachable nodes are visited. The result is a set of shortest distances from the starting node to all others and a tree structure representing the shortest paths. The algorithm's efficiency lies in its ability to guarantee that shorter paths are explored before longer ones, ensuring that the first instance a node is reached, it is via the shortest path. Furthermore, the time complexity of the BFS algorithm is O(V+E), where V is the number of vertices (nodes) and E is the number of edges in the graph. This is because each node and edge is processed once during the traversal. Additionally, alternative algorithms for

Algorithm 2: Breadth-first search

Input: Graph representing the company projection network and a starting node.

Output: a set of shortest distances from the starting node to all others and a tree structure representing the shortest paths.

```
// Initialize data structures
queue= empty_queue()
visited=set()
distance= {}
parent={}
// Enqueue the starting node
enqueue(queue, start)
visited.add(start)
distance[start]=0
parent[start]=None
// BFS traversal
while not is_empty(queue) do
    current_node=dequeue(queue)
    // Explore neighbors of the current node
    for neighbor in graph[current_node]: do
        if neighbor not in visited: then
            visited.add(neighbor)
            // Update distance and parent
            distance[neighbor] = distance[current_node] + 1
            parent[neighbor] = current_node
            // Enqueue the neighbor for further exploration
            enqueue(queue, neighbor)
```

finding the shortest paths include Dijkstra's algorithm and the Bellman-Ford algorithm. Dijkstra's algorithm is particularly efficient for graphs with non-negative edge weights and is well-suited for sparse graphs. On the other hand, the Bellman-Ford algorithm can handle graphs with negative edge weights and can detect negative cycles, but it may be slower than Dijkstra's algorithm, especially for sparse graphs. Both Dijkstra's and Bellman-Ford algorithms have a time complexity of $O(V^2)$ using an adjacency matrix or $O((V+E)logV)$ using an adjacency list, where V is the number of vertices and E is the number of edges [29]. The choice of the algorithm depends on the specific characteristics of the graph and the problem requirements. After we calculate the shortest path lengths between all nodes in the graph, we find the paths from each source node to the target node and we keep the shortest one. If the path does not exist, we assign it a value of 20. Finally, we save each source node and its shortest path length from a target node.

Now, we will consider a simplified example in order to get a better understanding of the metric. Looking at the Figure 5.1 we can see a bipartite graph on the left and its projection on the companies on the right. The nodes with red represents companies with a default event, while the green ones do not. Regarding the bipartite graph, the distance from 'c3' to 'c1' is 4 and from 'c2' to 'c1' is 2. On the other hand, on the projected network, the distance from 'c3' to 'c1' is 2 and from 'c2' to 'c1' is 1. From that we can conclude that in the bipartite graph the minimum distance that a company can have from another is 2 and in the projected network is 1. This happens because in the bipartite graph the companies belong to one of the two node sets and we cannot have a link between two nodes in the same set. On the contrary, in the projected network we have only one node set, so we can have a direct link from a company to another. In summary, the distance from default metric can be used in both bipartite and unipartite graphs with the difference that in the bipartite case the number of steps will be doubled than those in the unipartite case.

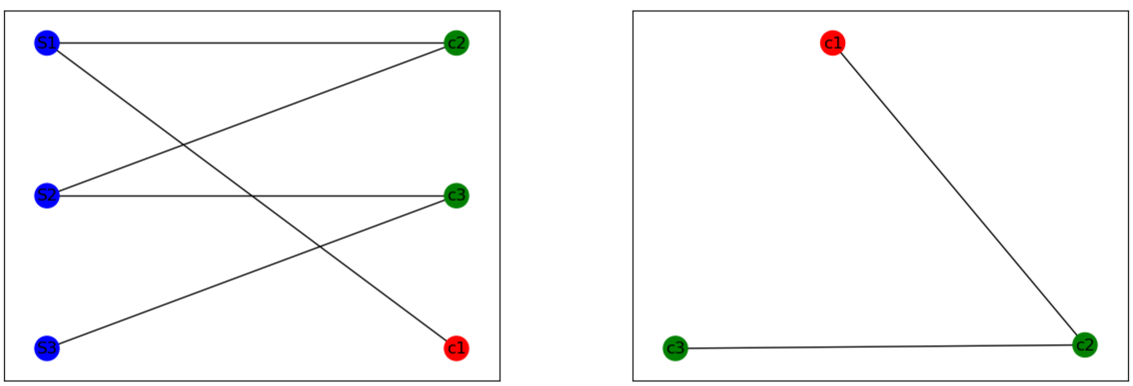

Figure 5.1: A bipartite graph on the left and its projection on the right.

In conclusion, the distance from default metric, although deceptively simple in its calculation, offers profound insights into credit risk assessment within complex fi-

 https://doi.org/10.26219/heal.aueb.6624

nancial networks. It acknowledges the interconnected nature of financial entities and provides a complete view of credit risk that extends beyond traditional financial indicators. By quantifying proximity to default events, this metric equips decision makers with a powerful tool for early warning, risk warning, and a deeper understanding of financial network dynamics.

5.2 Logistic Regression Models

In this section, we delve into constructing logistic regression models aimed at predicting the credit risk of each company. Our approach consists three distinct models, each with a singular focus on specific predictors. Initially, we use the network variables computed in Chapter 4, shedding light on their collective influence on credit risk. Subsequently, we focus to the individual metrics that we developed, namely Company Score and Distance from Default using them as exclusive predictors. These models not only help us understand the complex dynamics of credit risk prediction but also contribute valuable insights into the relative importance of network-based variables. The following sections detail the construction, results, and implications of each logistic regression model, providing a comprehensive understanding of their role in assessing credit risk.

5.2.1 Data Selection

As we mentioned in chapter 4, we will use the second dataset for evaluating our models. The dataset has quarterly financial information for every company spanning the period 2017 to 2022. For the evaluation of the models and our novel metrics, we focus on predicting credit risk. This is achieved by utilizing a binary variable, 'Target', where a value of 1 denotes a default event and 0 signifies the absence of such an event. As depicted in Figure 5.2, our dataset exhibits a temporal evolution of credit risk. The plot illustrates the number of default events observed over time, spanning from 187 to nearly 300 events. On average, approximately 32% of the companies each quarter experience a default event.

 https://doi.org/10.26219/heal.aueb.6624

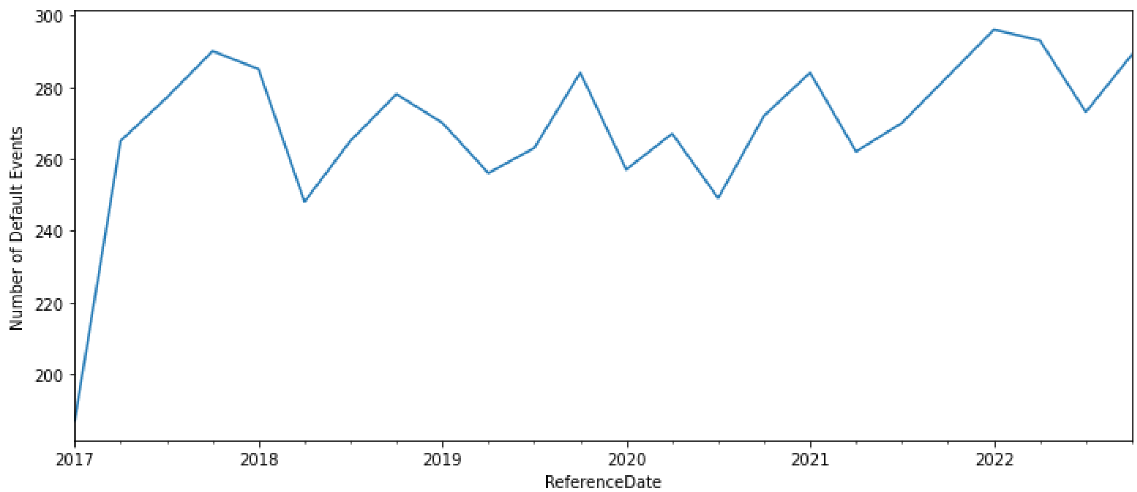

Figure 5.2: Plot of the number of default event versus time.

To capture the temporal dynamics inherent in credit risk assessment, we adopt a quarterly forecasting framework. Specifically, we divide the dataset into training and test sets. The training set encompasses information up to the year 2021 and serves as the basis for model training. In contrast, the test set consists data from each quarter of 2022 and is utilized to evaluate the performance of our models on a quarterly basis. This approach allows us to examine how our models perform in real-world scenarios, where timely and forward-looking predictions are important. Lastly, we will use a comprehensive set of metrics to assess the performance of the models. These measures include accuracy, precision, recall, and F1-score, providing an understanding of the model's effectiveness in different aspects. Additionally, we will use the Gini coefficient (also called Somers'D) as a valuable indicator of predictive power and discriminatory ability. The Gini coefficient, a well established performance metric for binary classification tasks, plays a pivotal role in our evaluation of the predictive capabilities of the models. It is calculated as: $Gini\_coefficient = 2AUC - 1$, where AUC represents the Area Under the Receiver Operating Characteristic (ROC) Curve. The ROC curve plots the true positive rate (sensitivity) against the false positive rate (1-specificity) for various threshold values, providing insight into the trade-off between true positive and false positive rates across different classification thresholds. AUC summarizes the overall performance of a binary classification model, with higher values indicating superior discrimination ability. Also, Gini coefficient takes values between -1 and 1, with 1 or -1 indicating that the independent variable perfectly predicts the dependent variable (1 when the relationship is positive and -1 when the relationship is negative). A value of 0 means that there is no relationship between the independent and dependent variable. Given our focus on credit risk, the Gini coefficient serves as a valuable tool for measure the models' effectiveness in distinguishing between creditworthy and non-creditworthy companies. Its insights enhance our ability to make informed decisions in the context of credit risk assessment.

5.2.2 Logistic Regression with Network Variables

In constructing our initial logistic regression model, we utilized degree, betweenness, closeness and eigenvector centrality, and the clustering coefficient as predictors and the Target variable as the response. These metrics were computed within the company-company network detailed in Chapter 4. The model was trained using the training dataset, and subsequently, we retained only those predictors that were statistically significant, with p-values below 0.05. As a result of this process, we can now write the final equation of our model:

$$\log\left(\frac{P(Target=1)}{1-P(Target=1)}\right) = -1.78 + 10.89 \cdot \text{closeness\_cent} - 8.08 \cdot \text{degree\_cent} \quad (5.3)$$

In Table 5.2, we can see for each predictor the standard error of its coefficient, the Z-statistic and the p-value. Notably, all predictors exhibit significance as evidenced by their p-values being less than 0.05. This statistical validation allows us to not only affirm the significance of each predictor but also to quantify the magnitude of their impact on the response variable. Also, from equation 5.3 we can conclude that the predictor variable closeness_cent has a positive coefficient of 10.89, which means that for each one-unit increase in the closeness centrality of a company in our network, the log-odds of the target variable being 1 (meaning a default event) increase by 10.89, holding the other variables constant. In practical terms, this suggests that higher closeness centrality is associated with a substantial increase in the likelihood of the company experienced a default event.

Predictors	**Coefficient**	**Std. Error**	**Z-Statistic**	**P-value**
Intercept	-1.78	0.04	-41.11	0
closeness_cent	10.89	0.44	24.89	0
degree_cent	-8.08	2.55	-3.16	0.002

Table 5.2: Summary of the model with the network variables as predictors

https://doi.org/10.26219/heal.aueb.6624

Next, we will evaluate our model using the test dataset and we will make predictions for each quarter in 2022. Looking at Table 5.3 we can see different measures for the evaluation of the model for each quarter. The model, demonstrates moderate performance across different quarters. While achieving an overall accuracy of approximately 0.51 to 0.52, the model excels in capturing true default instances (recall of 0.88 to 0.90). However, there is room for improvement in precision (0.40 to 0.42), indicating that when predicting defaults, the model occasionally generates false positives. The F1-Score, balancing precision and recall, falls between 0.55 and 0.57. The consistent Gini Coefficient values of around 0.20 to 0.21 suggest performance above random chance but leave potential for enhancing discriminatory ability.

Quarter	**Accuracy**	**Precision**	**Recall**	**F1-Score**	**Gini Coefficient**
$Q1$	0.52	0.42	0.89	0.57	0.20
$Q2$	0.51	0.41	0.88	0.56	0.20
$Q3$	0.51	0.40	0.90	0.55	0.21
$Q4$	0.52	0.42	0.90	0.57	0.21

Table 5.3: Different measures for the evaluation of logistic regression with network variables

The outcomes observed in our model evaluation may be affected by the static characteristics of the network variables compared to the dynamic evolution of the target variable over time. Recognizing this, there is a need to incorporate predictors that can capture the evolving dynamics of the network. To address this, we will use the two novel metrics, namely the company score and the distance from default, as predictors in our next phase. These metrics are designed to offer a better representation of the evolving network dynamics, aiming to improve the model's predictive capabilities in the challenging task of forecasting company defaults.

5.2.3 Logistic Regression with Company Score

In continuation of our previous methodology, we proceeded to develop a logistic regression model with the predictor being the company score. We used the equation 5.2 in order to calculate the scores for each company in our training set, which contains information from 2017 to 2021, and then we trained the model using the training set. The resulting equation for the model is as follows:

$$\log\left(\frac{P(Target=1)}{1-P(Target=1)}\right) = -4.03 + 8.62 \cdot \text{company\_score} \tag{5.4}$$

https://doi.org/10.26219/heal.aueb.6624

Looking at Table 5.4, we can see a summary of the model which includes the coefficient, standard error, Z-statistic and p-value for the intercept and the company score. We can see that the company score has a p-value equal to 0, which indicates that the company score is statistically significant. Also the coefficient of the company_score indicates that for every one-unit increase in the company score, the log-odds of the target being 1 increase by 8.62. In essence, the model suggests that higher company scores are associated with significantly higher odds of the company to have a default event.

Predictors	**Coefficient**	**Std. Error**	**Z-Statistic**	**P-value**
Intercept	-4.03	0.06	-66.15	0
company_score	8.62	0.14	63.03	0

Table 5.4: Summary of the model with the company score as predictor

As before, we will evaluate our model using the test dataset and we will make predictions for each quarter in 2022. The evaluation metrics for the logistic regression model, focusing on the company score, are summarized in the Table 5.5 for each quarter. The high accuracy scores across all quarters, ranging from 0.81 to 0.82, indicate the overall effectiveness of the model in correctly predicting the target variable. Precision scores, ranging from 0.73 to 0.75, highlight the model's ability to accurately identify positive instances. Recall scores, ranging from 0.69 to 0.75, demonstrate the model's capacity to capture a substantial portion of actual positive instances. The F1-Scores, ranging from 0.72 to 0.75, represent a balanced measure of precision and recall. The Gini Coefficient, ranging from 0.57 to 0.61, offers insights into the discriminatory power of the model. From these results, we can conclude the model's consistency and effectiveness across different quarters, providing a comprehensive understanding of its predictive performance in relation to the company score.

Quarter	**Accuracy**	**Precision**	**Recall**	**F1-Score**	**Gini Coefficient**
$Q1$	0.82	0.74	0.75	0.75	0.61
$Q2$	0.81	0.74	0.70	0.72	0.57
$Q3$	0.82	0.75	0.69	0.72	0.58
$Q4$	0.81	0.73	0.71	0.72	0.58

Table 5.5: Different measures for the evaluation of logistic regression with the company score

https://doi.org/10.26219/heal.aueb.6624

5.2.4 Logistic Regression with Distance from Default

The last model that we will make, uses as a predictor the distance from default metric where we introduced in section 5.1.2. Following the same procedure as previously, we trained the model utilizing the training dataset and we end up with the following equation for the model:

$$\log\left(\frac{P(Target = 1)}{1 - P(Target = 1)}\right) = -0.38 - 0.10 \cdot \text{dist\_from\_default} \tag{5.5}$$

The equation 5.5 reveals that the coefficient for the predictor dist_from_default is -0.10, indicating that with each one-unit increase, the log-odds of the target being 1 decrease by 0.10. The intercept, noted as -0.38, represents the log-odds of the target being 1 when dist_from_default is zero. Examining the summary in Table 5.6, we can verify the statistical significance of the predictors. Despite the significance of dist_from_default, it is obvious that its impact on the probability of default for a company is low. Lastly, we will use the test dataset in order to make prediction

Predictors	**Coefficient**	**Std. Error**	**Z-Statistic**	**P-value**
Intercept	-0.38	0.02	-20.19	0
dist_from_default	-0.10	0.01	-28.95	0

Table 5.6: Summary of the model with the distance from default as a predictor

for each quarter in 2022. Looking at Table 5.7, the accuracy values range from 0.51 to 0.55, indicating the overall proportion of correctly classified instances. Precision scores, ranging from 0.40 to 0.43, measure the accuracy of positive predictions. The recall values, varying from 0.85 to 0.92, signify the model's ability to capture true positive instances. F1-Score, combining precision and recall, ranges from 0.54 to 0.60, providing a balanced measure. The Gini Coefficient, ranging from 0.20 to 0.27, suggest performance above random chance.

Quarter	**Accuracy**	**Precision**	**Recall**	**F1-Score**	**Gini Coefficient**
$Q1$	0.55	0.43	0.91	0.59	0.25
$Q2$	0.53	0.42	0.90	0.58	0.23
$Q3$	0.51	0.40	0.85	0.54	0.20
$Q4$	0.55	0.43	0.92	0.60	0.27

Table 5.7: Different measures for the evaluation of logistic regression with the distance from default

https://doi.org/10.26219/heal.aueb.6624

5.3 Results

In this section, we present the outcomes of our study, drawing upon the methodologies and metrics introduced earlier. Initially, we explore the contributions of the two novel network metrics in enhancing our comprehension of network characteristics within shareholder-company networks and their application in credit risk assessment. Subsequently, we provide a concise summary of the results derived from the logistic regression models, aiming to identify the most effective model among them.

The company score metric emerges as a valuable tool for evaluating credit risk within a shareholder-company network. It provides a thorough assessment by taking into account the influence of shareholders and their investments in companies that have experienced default events. Tailored for application within bipartite networks, this metric assigns a numerical score ranging from 0 to 1, with a higher score signaling an increased likelihood of a company facing a default event. Importantly, the company score metric captures the dynamic nature of the network. As the composition of companies' shareholders evolves over time, the corresponding scores also adjust, allowing for a better understanding of credit risk that adapts to the changing dynamics of the network. In essence, the metric offers a straightforward representation of credit risk while also accounting for the evolving connections within the network.

The "Distance from Default" metric is a valuable tool for assessing credit risk within shareholder-company networks. It focuses on a fundamental question: how close or distant is a specific company from others in the network that have faced default events? By quantifying this proximity, the metric provides unique insights that go beyond traditional financial indicators, offering a comprehensive view that considers both individual financial health and network context. It recognizes the interconnected nature of companies, where financial events in one part of the network can impact connected entities. Despite its simple calculation, the metric proves to be a versatile and practical tool, aiding decision-makers in early warning, risk management, and gaining a deeper understanding of financial network dynamics. Applicable in both bipartite and unipartite networks, with values exclusively even in the bipartite case as we saw, a higher distance indicates a decreased likelihood of the company experiencing a default event.

The results presented in the Table 5.8 showcase the performance metrics of the three logistic regression models that we construct earlier, namely "Network Variables," "Company Score," and "Distance from Default," across different quarters. These metrics include Recall, F1-Score, and Gini coefficient. We focus on identifying the

Model	Quarter	Recall	F1-Score	Gini
Network Variables	Q1	0.89	0.57	0.20
	Q2	0.88	0.56	0.20
	Q3	0.90	0.55	0.21
	Q4	0.90	0.55	0.21
Company Score	Q1	0.75	0.75	0.61
	Q2	0.70	0.72	0.57
	Q3	0.69	0.72	0.58
	Q4	0.71	0.72	0.58
Distance from Default	Q1	0.91	0.59	0.25
	Q2	0.90	0.58	0.23
	Q3	0.85	0.54	0.20
	Q4	0.92	0.60	0.27

Table 5.8: Summary of performance metrics for the three models in each quarter

model that best predicts companies at risk of default, prioritizing metrics such as F1-score, recall, and the Gini coefficient, important measures in credit risk assessment. Notably, the model utilizing network variables demonstrates consistently high recall values ranging from 0.88 to 0.90, suggesting its proficiency in identifying true positives. However, its F1-score and Gini coefficient are comparatively lower. Conversely, the model employing the company score exhibits slightly lower recall values, ranging from 0.69 to 0.75, yet compensates with higher F1-scores ranging from 0.72 to 0.75 and Gini coefficient from 0.58 to 0.61. Lastly, the model leveraging the distance from default performs marginally better than the one with network variables.

Overall, our analysis underscores the value of the network metrics introduced, namely the "Company score" and "Distance from Default," in facilitating credit risk assessment. Through the models constructed, we observed that the distance from default metric exhibits slightly superior predictive capabilities compared to the traditional network variables model. Conversely, the model incorporating the company score metric surpassed both the distance from default and traditional network variables models in predictive accuracy. This suggests that the incorporation of the two novel network metrics significantly enhances predictive performance in credit risk assessment. Moreover, the robust predictive power of the company score metric highlights its potential as a key tool for financial institutions seeking to mitigate credit risk effectively. These findings underscore the importance of incorporating innovative metrics derived from network analysis into traditional credit risk assessment frameworks

https://doi.org/10.26219/heal.aueb.6624

Chapter 6

Conclusion

This concluding chapter serves as a reflection on the cumulative insights and contributions that have unfolded throughout this thesis. We commence by summarizing the key findings and outcomes draw from our comprehensive exploration of financial networks and the introduction of innovative network metrics. The pivotal role played by the company score and distance from default metrics in enhancing credit risk assessment becomes evident as we revisit our analysis. Lastly, we propose some future avenues that promise to deepen our understanding of financial networks and further refine credit risk assessment methodologies.

6.1 Summary of Findings

This section presents a comprehensive summary of the analysis conducted in chapter 4, where we explored the bipartite network and its projections, and in chapter 5, where we introduced two novel network metrics and built three models in order to assess their predictive capabilities in comparison with the traditional network variables.

In chapter 4, we explored the bipartite network, shedding light on its structural complexities. Calculations of degree and betweenness centrality provided valuable insights into the extent to which companies share common shareholders. The analysis of the projected networks allowed us to apply traditional network measures, enabling separate examination of the S-S network and the C-C network. Through the assessment of various network metrics, we identified the top 5 shareholders and companies which deserve special attention due to their significant influence within the network. These key entities exert substantial influence over the network's structure and dynamics, making them important points for further investigation and management.

https://doi.org/10.26219/heal.aueb.6624

In recognition of the limitations encountered during the analysis in chapter 4, where quantifying the risk of a company default remained difficult, we introduced two novel metrics: the 'Company Score' and 'Distance from Default'. The company score assigns a value between 0 and 1 to each company, with higher scores indicating an elevated probability of default. On the other hand, the distance from default metric calculates the shortest path lengths from companies without default events to those with such events, with higher distances signifying a reduced susceptibility to the default of other companies. Utilizing the dataset introduced in Chapter 4, we constructed three predictive models, leveraging information up to 2021 as the training set. Our objective was to assess the predictive capabilities of traditional network metrics, the company score, and the distance from default for each quarter in 2022. Notably, the model utilizing traditional network variables exhibited the poorest performance. This can be attributed to the static nature of the predictors, which failed to adapt to the changing dependent variable over time. In contrast, the introduced network metrics possess inherent time-dependency, enabling them to better capture the evolving dynamics of the network. While the distance from default yielded marginally improved predictions compared to traditional network variables, its performance remained sub-optimal. Remarkably, the company score emerged as the standout performer, demonstrating significantly superior predictive power compared to its counterparts.

Finally, this synthesis of findings underscores the potential of the two new network metrics to advance our understanding of credit risk within financial networks. The company score and distance from default metrics, hold promise as valuable tools in credit risk assessment, offering fresh perspective and avenues for further research in the dynamic realm of financial networks.

6.2 Future Work

Within the domain of financial networks and credit risk assessment, opportunities for advancing our comprehension and refining predictive models are plenty. This section outlines prospective avenues for future research, each aimed at enhancing our understanding of financial networks and their implications for credit risk assessment.

First, the predictive capabilities of the introduced metrics could be assessed in multivariate models. Combining these metrics with other relevant financial variables may enhance prediction accuracy. Furthermore, future research can explore the concept

https://doi.org/10.26219/heal.aueb.6624

of evolving networks. Allowing nodes and links to evolve over time can facilitate the analysis of temporal dynamics within financial networks. This temporal evolution analysis can uncover patterns, trends, and transitions in credit risk profiles over time. Moreover, it may be necessary to adapt network metrics to accommodate time-dependent attributes, thus offering a more comprehensive understanding of how financial networks evolve. Lastly, expanding the scope of this study to encompass diverse financial ecosystems and industries may yield insights into sector-specific credit risk dynamics. These avenues represent promising directions for future research, with the potential to refine and extend the application of network-based metrics in the domain of credit risk assessment.

https://doi.org/10.26219/heal.aueb.6624

Bibliography

[1] M. E. J. Newman. *Networks: an introduction*. Oxford University Press, Oxford; New York, 2010.

[2] Phillip Bonacich. Power and centrality: A family of measures. *American Journal of Sociology*, 92(5):1170–1182, 1987.

[3] Leo Katz. A new status index derived from sociometric analysis. *Psychometrika*, 18, 1953.

[4] Lawrence Page and Sergey Brin. The anatomy of a large-scale hypertextual web search engine. *Computer Networks*, 30, 1998.

[5] Linton C. Freeman. A set of measures of centrality based on betweenness. *Sociometry*, 40(1):35–41, 1977.

[6] Michael Boss, Helmut Elsinger, Martin Summer, and Stefan Thurner. The network topology of the interbank market *, 2003.

[7] Stefano Battiston, Michelangelo Puliga, Rahul Kaushik, Paolo Tasca, and Guido Caldarelli. Debtrank: Too central to fail? financial networks, the fed and systemic risk. *Scientific Reports*, 2, 2012.

[8] AOYAMA Hideaki, Stefano BATTISTON, and FUJIWARA Yoshi. DebtRank Analysis of the Japanese Credit Network. Discussion papers 13087, Research Institute of Economy, Trade and Industry (RIETI), October 2013.

[9] Stefan Thurner and Sebastian Poledna. Debtrank-transparency: Controlling systemic risk in financial networks. 1 2013.

[10] Marco Bardoscia, Stefano Battiston, Fabio Caccioli, and Guido Caldarelli. Debtrank: A microscopic foundation for shock propagation. *PLoS ONE*, 10, 6 2015.

https://doi.org/10.26219/heal.aueb.6624

[11] Thiago Christiano Silva, Sergio Rubens Stancato Souza, and Benjamin Miranda Tabak. Monitoring vulnerability and impact diffusion in financial networks. *Journal of Economic Dynamics and Control*, 76:109–135, 3 2017.

[12] Miia Chabot and Jean Louis Bertrand. Complexity, interconnectedness and stability: New perspectives applied to the european banking system. *Journal of Business Research*, 129:784–800, 5 2021.

[13] Walter Cuba, Anahi Rodriguez-Martinez, Diego A. Chavez, Fabio Caccioli, and Serafin Martinez-Jaramillo. A network characterization of the interbank exposures in peru. *Latin American Journal of Central Banking*, 2:100035, 9 2021.

[14] Isela-Elizabeth Téllez-León, Serafín Martínez-Jaramillo, Luis O. L. Escobar-Farfán, and Ronald Hochreiter. How are network centrality metrics related to interest rates in the Mexican secured and unsecured interbank markets? *Journal of Financial Stability*, 55(C), 2021.

[15] Michelle Girvan and M E J Newman. Community structure in social and biological networks, 2001.

[16] M. Newman. Newman, m.e.j.: Detecting community structure in networks. eur. phys. j. b 38(2), 321-330. *Physics of Condensed Matter*, 38:321–330, 03 2004.

[17] E. A. Leicht and M. E. J. Newman. Community structure in directed networks. 9 2007.

[18] Vincent D. Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. Fast unfolding of communities in large networks. 3 2008.

[19] M. Rosvall and C. T. Bergstrom. Maps of random walks on complex networks reveal community structure. 7 2007.

[20] Stefania Vitali and Stefano Battiston. The community structure of the global corporate network e104655. *PLoS ONE*, 9, 8 2014.

[21] Yajing Huang and Feng Chen. Community structure and systemic risk of bank correlation networks based on the u.s. financial crisis in 2008. *Algorithms*, 14, 2021.

[22] Wataru Souma, Yoshi Fujiwara, and Hideaki Aoyama. Shareholding networks in japan, 2005.

https://doi.org/10.26219/heal.aueb.6624

[23] MARCO D'ERRICO, Rosanna Grassi, Silvana Stefani, and Anna Torriero. *Shareholding Networks and Centrality: An Application to the Italian Financial Market*, volume 613, pages 215–228. 01 2009.

[24] Huajiao Li, Wei Fang, Haizhong An, and Lili Yan. The shareholding similarity of the shareholders of the worldwide listed energy companies based on a two-mode primitive network and a one-mode derivative holding-based network. *Physica A: Statistical Mechanics and its Applications*, 415:525–532, 12 2014.

[25] Qing Yao, Tim S. Evans, and Kim Christensen. How the network properties of shareholders vary with investor type and country. *PLoS ONE*, 14, 8 2019.

[26] Jingjing Long, Cuiqing Jiang, Stanko Dimitrov, and Zhao Wang. Clues from networks: quantifying relational risk for credit risk evaluation of smes. *Financial Innovation*, 8, 12 2022.

[27] Ruijie Liu and Yajing Huang. Structural analysis of projected networks of shareholders and stocks based on the data of large shareholders' shareholding in china's stocks. *Mathematics*, 11, 3 2023.

[28] Nathan R. Todd, Jaclyn D. Houston, and Rachael L. Suffrin. Applying affiliation social network analysis to understand interfaith groups. *Psychosocial Intervention*, 24(3):147–154, 2015. Network Analysis for Social and Community Interventions[Análisis de redes sociales e intervención comunitaria].

[29] Thomas H. Cormen, Charles E. Leiserson, Ronald L. Rivest, and Clifford Stein. *Introduction to Algorithms*. The MIT Press, 2nd edition, 2001.