Stockholm University

Graphical Models

Mathematical Foundation and Statistical Analysis

Elena Farahbakhsh Touli

Doctoral Thesis in Computational Mathematics at Stockholm University, Sweden 2024

Graphical Models

Mathematical Foundation and Statistical Analysis

Elena Farahbakhsh Touli

Academic dissertation for the Degree of Doctor of Philosophy in Computational Mathematics at Stockholm University to be publicly defended on Thursday 22 August 2024 at 13.00 in Cramérrummet (mötesrum 12), hus 1, Albano, Albanovägen 28.

Abstract

Working on different problems related to graph theory combined with statistics is the main purpose of this thesis.

In paper I, we worked on the distance between trees and defined another definition for the interleaving distance that was already introduced for determining the distance between merge trees. The new definition was based on only one map from one of the trees to another one. Therefore, we could gain fixed-parameter tractable algorithms for finding the interleaving distance between merge trees with some conditions.

In paper II, we worked on the clustering coefficient of the networks. The clustering coefficient indicates the tendency of the vertices of the network to form a triangle. We introduced another clustering coefficient, which we called it Relative clustering coefficient. Finally, the importance of the relative clustering coefficient.

In paper III, we worked on the financial relationship between companies in Sweden, and we used two methods (the Pearson correlation coefficient (PCC) and the generalized variance decomposition (GVD). We then ap- plied these methods to the financial data consisting of the daily returns on the 28 stocks included in the computation of the OMX index (the index of the Swedish capital market).

Gaussian Graphical Model is the main subject of paper IV. In this paper, we consider three types of precision matrices, and corresponding to each type of precision matrix, we develop the exact test theory. Finally, the new approaches are compared to the benchmark method via an extensive simulation study.

Stockholm 2024
http://urn.kb.se/resolve?urn=urn:nbn:se:su:diva-229153

ISBN 978-91-8014-833-7
ISBN 978-91-8014-834-4

Department of Mathematics

Stockholm University, 106 91 Stockholm

GRAPHICAL MODELS

Elena Farahbakhsh Touli

Stockholm
University

Graphical Models

Mathematical Foundation and Statistical Analysis

Elena Farahbakhsh Touli

ISBN print 978-91-8014-833-7
ISBN PDF 978-91-8014-834-4

Printed in Sweden by Universitetsservice US-AB, Stockholm 2024

To my father in heaven.

Acknowledgements

During my studies in high school, I was highly encouraged by some of my teachers to study in the field of Mathematics. After the National exam, I was accepted at KNT University in Tehran, Iran.
In the first semester, I had a lovely teacher Prof. Hassan Haghighi and he encouraged me a lot and told me: ”Elena! If you want to succeed in this field, you should improve your English language”. Also, he helped me to write a basic program by Maple to find all the prime numbers less than 100. However, after one semester as some of my friends encouraged me to become a medical doctor and they told me that my style looked like a medical student’s style, I decided to study Medicine.
Nonetheless, after taking the national exam my score was 250 and the last person accepted in that university was 200. I could get a better grade, however, there were three questions that I had doubts about, and one of them was about the heart of spiders, just after the exam, I understood that I did it wrong. At that time I understood that learning about mathematical functions is easier for me than insects. So, I came back to continue my studies in the field of Mathematics, but with a stronger motivation to get my Ph.D. in Mathematics. At KNTU, I took some courses in Computer Science, and with a lot of effort, I could get my bachelor's degree after seven semesters (In Iran getting a bachelor's degree takes 8 semesters) with the first rank among all the students in the field of Mathematics to be accepted at the Sharif University of Technology without taking the national exam.
I want to thank one of my friends Mohammad Javad Salari Seddigh whom I met at KNTU and later he helped me to open a big door so I could get a chance to study abroad. Also, Sajedeh Farzaneh Doost who is my cousin, and my aunt Nahid Malek Moghaddam who is a teacher whom I spent the most time with during my studies in Tehran. At Sharif, I met some very

knowledgeable, very kind, and helpful teachers like Prof. Nezam Mahdavi-Amiri who was my advisor during my master's study, and Prof. Mahmoud Hesaaraki. In 2016 I attended The Ohio State University to do research under the supervision of Prof. Yusu Wang. Working under her supervision and also taking a course by Prof. Anastasios Sidiropoulos were honors for me and they helped me a lot to broaden my knowledge in Computer Science and Graph Theory. At The Ohio State University, I had some nice colleagues such as Dr. Alfred Rossi, Dr. Minghao Tian, and Dr. Dingkang Wang.
They were very supportive and my best friends during my work at OSU. After some years, I came to Sweden to pursue my Ph.D. at Stockholm University to become a doctor, not a medical doctor. Here, I could meet my advisor Prof. Olha Bodnar who helped me during my Ph.D. and supported me.
She was so understanding and listening. She helped me and taught me statistics and I learned a lot during the course that I had with her at Örebro University. I can say that she taught me not only statistics but also life lessons.
Thanks, Olha! Here at SU, I also met three lovely and supportive people who helped me and supported me a lot and they are Prof. Joanna Tyrcha, Neshat Lindberg, and Dr. Abhishek Pal Majumder. Moreover, the last year of my Ph.D. I was employed by Högskolan i Gävle at the Department of Occupational Health, Psychology, and Sports Sciences. It was the best part of my PhD; all the people were very nice, helpful, and friendly, and I cannot choose some of them to say thanks.
Finally, I want to say thanks to my father in heaven that his life did not allow him to be here, and celebrate this day with me and my mother for their immeasurable support and love. For their encouragement during all of my life to study better and work harder. Also, thanks to Aref, for his support during my life in Sweden. And, thanks to my lovely sister. She was always for me to talk to me and support me without judging me not only when I did the right things, but also when I did the wrong ones.
All these people whom I mentioned their names here have continuously directly or indirectly helped me get here. I will never forget them, even though, some of them may never read this page. And, my goals, wishes, and hard work will not end by getting the Ph.D. degree.

Contents

Abstract 6
List of Papers 7
Acknowledgements 10
1 Introduction 12
2 Summary of the Papers 24
3 Conclusion and Further Work 31
4 Svensk Sammanfattning 32

Bibliography **36**

Abstract

Working on different problems related to graph theory combined with statistics is the main purpose of this thesis.

In paper I, we worked on the distance between trees and defined another definition for the interleaving distance that was already introduced for determining the distance between merge trees. The new definition was based on only one map from one of the trees to another one. Therefore, we could gain fixed-parameter tractable algorithms for finding the interleaving distance between merge trees with some conditions.

In paper II, we worked on the clustering coefficient of the networks. The clustering coefficient indicates the tendency of the vertices of the network to form a triangle. We introduced another clustering coefficient, which we called it Relative clustering coefficient. Finally, the importance of the relative clustering coefficient.

In paper III, we worked on the financial relationship between companies in Sweden, and we used two methods (the Pearson correlation coefficient (PCC) and the generalized variance decomposition (GVD)). We then applied these methods to the financial data consisting of the daily returns on the 28 stocks included in the computation of the OMX index (the index of the Swedish capital market).

Gaussian Graphical Model is the main subject of paper IV. In this paper, we consider three types of precision matrices, and corresponding to each type of precision matrix, we develop the exact test theory. Finally, the new approaches are compared to the benchmark method via an extensive simulation study.

List of Papers

The thesis consists of four papers:

I Elena Farahbakhsh Touli, Yusu Wang (2022), FPT-Algorithms for Computing Gromov-Hausdorff and Interleaving Distances between Trees, *Journal of Computational Geometry* 13(1): 89-124.

II Elena Farahbakhsh Touli, Oscar Lindberg (2022), Relative Clustering Coefficient. *Journal of Algorithms and Computation* 54(1): 99-108.

III Elena Farahbakhsh Touli, Hoang Nguyen, Olha Bodnar (2024), Monitoring the Dynamic Networks of Stock Returns with an Application to the Swedish Stock Market. To appear in *Computational Economics*.

IV Olha Bodnar, Elena Farahbakhsh Touli (2023), Exact Test Theory in Gaussian Graphical Models, *Journal of Multivariate Analysis* 196: 105185.

Author's contributions: In the first paper, my contribution was as follows: Conceptualization, Methodology, Analysis, Writing - Original Draft, Writing - Review and Editing, and Visualization. In the second paper, I did Conceptualization, Methodology, Validation, Resources, Writing - Original Draft, Writing - Review and Editing, Supervision, and Visualization. My contribution to the third paper was Methodology, Software, Formal Analysis, Data Curation, Writing - Original Draft, Writing - Review and Editing, and Visualization. My contribution to the fourth paper was Conceptualization, Methodology, Resources, Visualization, and Writing - Original Draft.

List of Figures

1 Two similar trees with a large alignment distance. ε is a very small number. 18
2 Two different trees T_1 and T_2 with a small Hausdorff distance. 19
3 Three types of graphs: left-hand-side plot corresponds to the AR(1) structure of the precision matrix, middle plot the block-diagonal structure of the precision matrix, right-hand-side plot factor structure. From "Exact Test Theory in Gaussian Graphical Models", by Olha Bodnar, Elena Farahbakhsh Touli, 2023, Journal of Multivariate Analysis 196: 105185. Copyright 2024 by Elena Farahbakhsh Touli with permission. 30

Acknowledgements

During my studies in high school, I was highly encouraged by some of my teachers to study in the field of Mathematics. After the National exam, I was accepted at KNT University in Tehran, Iran. In the first semester, I had a lovely teacher Prof. Hassan Haghighi and he encouraged me a lot and told me: "Elena! If you want to succeed in this field, you should improve your English language". Also, he helped me to write a basic program by Maple to find all the prime numbers less than 100. However, after one semester as some of my friends encouraged me to become a medical doctor and they told me that my style looked like a medical student's style, I decided to study Medicine. Nonetheless, after taking the national exam my score was 250 and the last person accepted in that university was 200. I could get a better grade, however, there were three questions that I had doubts about, and one of them was about the heart of spiders, just after the exam, I understood that I did it wrong. At that time I understood that learning about mathematical functions is easier for me than insects. So, I came back to continue my studies in the field of Mathematics, but with a stronger motivation to get my Ph.D. in Mathematics. At KNTU, I took some courses in Computer Science, and with a lot of effort, I could get my bachelor's degree after seven semesters (In Iran getting a bachelor's degree takes 8 semesters) with the first rank among all the students in the field of Mathematics to be accepted at the Sharif University of Technology without taking the national exam. I want to thank one of my friends Mohammad Javad Salari Seddigh whom I met at KNTU and later he helped me to open a big door so I could get a chance to study abroad. Also, Sajedeh Farzaneh Doost who is my cousin, and my aunt Nahid Malek Moghaddam who is a teacher whom I spent the most time with during my studies in Tehran. At Sharif, I met some very knowledgeable, very kind, and helpful teachers like Prof. Nezam Mahdavi-Amiri who was my advisor during my master's study, and Prof. Mahmoud Hesaaraki. In 2016 I attended The Ohio State University to do research under the supervision of Prof. Yusu Wang. Working under her supervision and also taking a course by Prof. Anastasios Sidiropoulos were honors for me and they helped me a lot to broaden my knowledge in Computer Science and Graph Theory. At The Ohio State University, I had some nice colleagues such as Dr. Alfred Rossi, Dr. Minghao Tian, and Dr. Dingkang Wang. They were very supportive and my best friends during my work at OSU. After some years, I came to Sweden to pursue my Ph.D. at Stockholm University to become a doctor, not a medical doctor. Here, I could meet my advisor Prof. Olha Bodnar who helped me during my Ph.D. and supported me. She was so

understanding and listening. She helped me and taught me statistics and I learned a lot during the course that I had with her at Örebro University. I can say that she taught me not only statistics but also life lessons. Thanks, Olha! Here at SU, I also met three lovely and supportive people who helped me and supported me a lot and they are Prof. Joanna Tyrcha, Neshat Lindberg, and Dr. Abhishek Pal Majumder. Moreover, the last year of my Ph.D. I was employed by Högskolan i Gävle at the Department of Occupational Health, Psychology, and Sports Sciences. It was the best part of my PhD; all the people were very nice, helpful, and friendly, and I cannot choose some of them to say thanks.

Finally, I want to say thanks to my father in heaven that his life did not allow him to be here, and celebrate this day with me and my mother for their immeasurable support and love. For their encouragement during all of my life to study better and work harder. Also, thanks to Aref, for his support during my life in Sweden. And, thanks to my lovely sister. She was always for me to talk to me and support me without judging me not only when I did the right things, but also when I did the wrong ones.

All these people whom I mentioned their names here have continuously directly or indirectly helped me get here. I will never forget them, even though, some of them may never read this page. And, my goals, wishes, and hard work will not end by getting the Ph.D. degree.

1 Introduction

Graph theory and statistical analysis play an important role in understanding the structure and dynamics of complex systems. In this thesis, we address the complex interplay between these two fields and focus on key concepts such as graphs, distance between trees, clustering coefficients, statistical inference theory, and graphical models and their applications in various fields such as economics and finance.

Graphs provide a versatile framework for representing the relationships between objects. In the domain of computational geometry we address the concept of distance between trees, a measure used to quantify the difference or similarity between trees.

The clustering coefficient measures the tendency of nodes in a graph to form local clusters or cliques. We discuss how this measure can be extended to a new definition for finding the clustering coefficient in the network.

Statistical analysis serves as the cornerstone of financial data discovery and interpretation. From descriptive statistics to inferential methods, statistics provide powerful tools for summarizing, analyzing, and drawing meaningful conclusions from financial data sets. We address the application of statistical techniques in network analysis and emphasize their role in hypothesis testing, data-driven discovery, and predictive modeling.

By integrating concepts from graph theory and statistical analysis, we gain deeper insight into the structure and function of networks.

This chapter introduces some basic definitions of mathematics and computer science used in this thesis. In sections 1.1, 1.2, and 1.3, we discuss several theoretical facts related to the computational complexity and the distance between trees that were used in the first paper. Section 1.4 illustrates several useful definitions of time series analysis and finance that were used in the third paper. Finally, section 1.5 summarizes the papers included in the thesis on which I worked during my Ph.D. study at Stockholm University.

1.1 NP-Completeness

In this section, the sets of problems in P and NP are defined, and we used them in the first paper. The main references for this section are [GJ79; Cor+01].

A problem is said to be a *decision problem* if its answer is 'yes' or 'no'. For example, for a given ε whether the distance between two trees is less than ε or not? Another type of problems is considered as an optimization

problem. For example, for two given trees, what is the distance between them? We can translate this problem as what the minimum ε is such that the distance between two trees is less than or equal to ε.

Definition 1. *(P)*
The set of decision problems that there exist polynomial time algorithms to solve them is called P.

For example, one can ask a question whether there exists a pair of vertices in a graph $G = (V, E)$ that the shortest distance between them is ε. To answer this question, we need to find the shortest distance between all the pairs of nodes and then check if there is at least one with the length of ε. This results in a polynomial time algorithm. See [Cor+01] for more information.

Definition 2. *(NP)*
The set of decision problems that we can verify in a polynomial time is called NP.

But what does it mean to say that the problem is verifiable in polynomial time? It means that there is a polynomial time algorithm that if we give the solution as the input, in a polynomial time it will return "yes".

For example, the 3-colorable problem in graphs is NP. This problem is formulated as follows: For a given graph $G = (V, E)$ and three colors C_1, C_2, and C_3, if we can indicate each node by a color such that an edge does not connect two nodes of the same color. This problem is NP and we can verify it in a polynomial time. Because, if we have the solution to this problem for this graph, then we can check in a polynomial time if the same colors are connected or not.

Every problem in P is also polynomial time verifiable, therefore, $P \subset NP$.

Definition 3. *(NP-complete)*
A decision problem is called NP-complete *if it is in NP and any problem in NP can be reduced to it in polynomial time.*

For an example, 3-coloring is NP-complete (see, e.g., [Cor+01] for the proof). When we say that 3-coloring is NP-complete, then we do not claim that there is no known algorithm for solving it. There is, of course, a non-polynomial time algorithm for finding 3-coloring in any graph but we do not know whether there is any *polynomial* time algorithm or not. One of the algorithms for 3-coloring is just to consider all the possibilities.

We say that a problem (say A) is reducible to another problem (say B) in a polynomial time if we can convert problem A into B in a polynomial

time and we write $A \leq_p B$. And it can be said in slang that the problem B is at least as hard as problem A. Also, if there is any polynomial time algorithm for solving problem B, so there is one for A as well.

Note 1. *To prove that a decision problem (say A) is in NP-complete we just need to prove that it is polynomial time verifiable and also find one problem in NP-complete (say B) and reduce B to A. Because we already know that any problem in NP-complete can be reduced to B in a polynomial time, therefore if we can reduce B to A in a polynomial time, we can reduce all the polynomials in NP to A in a polynomial time.*

For more illustration, let us consider an example. In the following, we will prove that $P1$ which is defined in the following is NP-complete.
P1: Is the set of solutions of a system of equations of degree at most three empty or not?

Theorem 1. *([Gal13]) 3-coloring can be reduced to P1 in polynomial time.*

Proof. First, we need to prove that this problem is verifiable in polynomial time. To do so, for a given solution of the set, we only need to see if they are the solution of each polynomial in the set or not. Then, if we consider a graph with the nodes $V_1, V_2, ..., V_n$, and three colorings denoted by $\{-1, 0, 1\}$ which corresponds to for example $\{red, blue, green\}$ for each node, then we have three possibilities that each node V_i can be red, blue or green. So, it means that x_i which is the color of node V_i can be -1, 0, or 1. Therefore, for each node, we have one equation which is $(x_i - 1)(x_i)(x_i + 1) = x_i^3 - x_i = 0$. Other than this set of equations, we have the condition that no two connected nodes (with an edge between them) have the same color. In other words, for any edge e_{ij} if $x_i = 1$, x_j cannot be 1, and so will be 0 or -1. Therefore, for each edge e_{ij} we have that $x_i^2 + x_i x_j + x_j^2 = 1$. Hence, we find a system of polynomials of the degree of at most three with n (the number of nodes of the graph) variables and $n + m$ polynomials where m is the number of edges. Therefore, in a polynomial time, we reduced the problem of 3-coloring to the existence of a solution of a system of polynomials of degree at most three. Thus, if there is a polynomial time algorithm for P1, then we could solve 3-coloring in a polynomial time. □

It has not been proven if P=NP or not. However, researchers tend to accept that $P \neq NP$. Therefore, they try to solve problems with different methods.

Definition 4. *(NP-hard)*
A decision problem is called NP-hard *if any problem in NP can be reduced to it in a polynomial time.*

1.1.1 Approximation Algorithms

There are some optimization problems that their decision problems are NP-complete, but they are very important that we look for a polynomial time algorithm that can give us a near-optimal solution. An approximation algorithm for any input returns a value C such that for the maximization problem (minimization problem) we have that $0 < C < C^*$ $(0 < C^* < C)$ where C^* is the optimal value.

We have two types of approximation algorithms that are defined as follows:

Definition 5. *(Absolute approximation)*
If for any instance of an optimization problem, C is the value that the approximation algorithm returns and C^* is the optimal value we have that

$$|C - C^*| \leq \varepsilon.$$

Definition 6. *(Relative approximation)*
If for any instance of an optimization problem, C is the value that the approximation algorithm returns and C^* is the optimal value we have that

$$\max\{\frac{C}{C^*}, \frac{C^*}{C}\} \leq 1 + \varepsilon.$$

In this case, we say that we have a $(1+\varepsilon)$-approximation algorithm for the problem.

Now we define another type of problem which is called Max SNP-hard problems as follows: (For more illustration see [ZJ94a; Aro+92])

Definition 7. *(PTAS)*
A problem has a polynomial time approximation schema (PTAS) if, for any $\varepsilon > 0$, there exists a polynomial time $(1+\varepsilon)$-approximation algorithm.

Definition 8. *(MAX SNP-hard)*
A problem is called to be MAX SNP-hard (Strictly NP-hard) if it does not have a PTAS unless $P = NP$.

1.2 Graph Theory: Definitions and Basic Properties

In mathematics and computer science, the term *graph* is one of the vast subjects. In graph theory, there are a lot of open problems and topics

that require further investigation. Therefore, it attracts the attention of different scientists around the world.

A graph $G = (V, E)$ is a structure with a set of vertices denoted by V and a set of edges E that indicates the relation between the vertices, which is shown by an edge between pairs of vertices (For more illustration see [Cor+01]). Any graph is represented by using a matrix, called *Adjacency Matrix* defined as follows:

Definition 9. *(Adjacency Matrix) [Cor+01]*
For any unweighted undirected graph $G = (V, E)$, there exists an n by n symmetric matrix A, called Adjacency Matrix*, that represents the graph where n is the number of vertices in the graph. Any element A_{ij} of the matrix A is* 1 *if there is an edge between the i-th vertex and the j-th vertex.*

Let w_{ij} be the weight of the edge e_{ij}. If the graph is directed which means that there exist $i, j \in \{1, ..., n\}$ such that $w_{ij} \neq w_{ji}$, then the element A_{ij} of the adjacency matrix A is w_{ij} and the adjacency matrix is not symmetric [Cor+01].

There is another matrix that is defined in graph theory which is called *Distance Matrix*. The definition of distance matrix is as follows:

Definition 10. *(Distance Matrix) [Cor+01]*
For a weighted undirected graph, the distance matrix is defined as a symmetric matrix D that each element D_{ij} of the matrix indicates the shortest distance from V_i to V_j.

We can define the distance matrix for unweighted graphs as well, in which each element D_{ij} indicates the number of edges of the shortest path between i-th vertex and the j-th vertex.

The distance matrix is essential for finding the center of a graph. In finance, the center of a network or a graph that we construct from the financial data is very important. By finding the center of the graph, we can find which company has the largest effect on other companies. For example, if the company that is in the center of the graph becomes bankrupt, then all the other companies in the network will be affected by this bankruptcy in the shortest time. The definition of the *center* is the following:

Definition 11. *(Center) [WF94]*
In a graph, the center is a vertex that has the shortest distance to all the other vertices of the graph.

Finding the center in the graph is described by the following procedure: For each vertex V_i, we find the shortest distance of V_i to all the other

vertices and then we choose the largest number between them, which is called max-distance of vertex i. Finally, the center of the graph is the vertex that has the minimum max-distance.

1.3 Trees and their Similarity

During my Ph.D. I have been interested in a special kind of graph that is called *tree*, especially I have been interested in *merge trees*. A tree is a connected graph (i.e., for any pairs of nodes in the graph, there is at least a path that connects these two nodes) without cycles, while the merge tree is defined in the following by defining the metric tree:

Definition 12. *(Metric Trees)* [MBW13]
A metric space $(|T|, d)$ is a metric tree where $|T|$ is the underlying space and d is the shortest distance between two points in T. The shortest distance between two nodes v_i and v_j in an undirected graph is the shortest distance of the path between v_i and v_j among all the paths that connect v_i and v_j.

We can embed any tree to $\mathbb{R}^2$ or $\mathbb{R}^3$ and the distance between any two nodes is the shortest distance between the nodes.

Definition 13. *(Merge Tree)* [MBW13]
A merge tree is a rooted tree that is embedded into R^2 or R^3 with a continuous function $f : |T| \to R$ which is defined on it. The function f decreases when we consider points from the root of the tree to each leaf.

Comparison of trees is used in the field of medicine and for comparing the phylogeny trees and in many other areas [MBW13; Smi20; Bil05]. In the thesis, we are interested in finding an efficient algorithm to compare trees in a way that the relation between children and parents does not affect the distance a lot. For more illustration look at Figure 1.

1.3.1 Distance between Trees

The distance between two trees is the main subject of the first paper. There are a few ways to calculate the similarity between trees. However, either they gave us little information about the similarity between trees, or they were not efficient in calculating them. Here we look at some distances between two trees.

The first distance between two trees that we talk about is the *tree edit distance*. The tree edit distance is the most common distance for measuring the dissimilarity between two trees.

Definition 14. *(Tree Edit Distance)* [Bil05]
The tree edit distance is defined as the optimal cost for converting one tree to another one by using three operations renaming, adding, and removing.

Another distance that we consider for finding the similarity between trees is *tree alignment distance.*

Definition 15. *(Tree Alignment Distance)* [Bil05]
The tree alignment distance is defined as the optimal cost for renaming the nodes after adding nodes and edges to both trees to reach the same structure.

It has been proven in [JWZ95; ZJ94b] that calculating both distances between trees is MAX SNP-hard to compute for arbitrary trees. However, there are polynomial time algorithms for computing the tree edit distance between two labeled ordered trees[1] with a bounded depth and also there is a polynomial time algorithm for computing the tree alignment distance between any two labeled ordered trees with bounded degrees of nodes [JWZ95]. In practice, it is difficult to bound the depth of the trees, but it is very common to consider trees with bounded degrees. Although there is a polynomial time algorithm to calculate the tree alignment distance between labeled ordered trees with bounded degrees, we are not interested in the tree alignment distance because the tree alignment distance depends highly on the relation between children and parents. For example, if we look at trees in Figure 1, then we see that two trees are very similar to each other, but the alignment distance between them can be arbitrarily large.

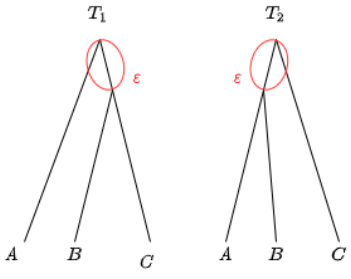

Figure 1: Two similar trees with a large alignment distance. ε is a very small number.

Some other distances have been defined between metric trees.

[1]Ordered tree is a tree that there is an order between the children of nodes.

Definition 16. *(Hausdorff Distance)* [ES97]
Let M be a metric space with distance d. Also, let $X, Y \subseteq (M, d)$ be two subsets of the metric space (M, d). The Hausdorff distance between X and Y is defined as follows:

$$d_H(X, Y) = \max\{\sup_{x \in X} \inf_{y \in Y} d(x, y), \sup_{y \in Y} \inf_{x \in X} d(x, y)\}.$$

We cannot use the Hausdorff distance to compare two metric trees and the reason is that for two different trees, it can give us a small number as illustrated in Figure 2.

Figure 2: Two different trees T_1 and T_2 with a small Hausdorff distance.

Another distance that we can use for comparing trees is the Gromov-Hausdorff distance. The Gromov-Hausdorff distance is defined between any two metric spaces and it is given as follows:

Definition 17. *(Gromov-Hausdorff Distance)* [Mem07; K A+18; TW22]
Let $\mathbb{X} = (X, d_X)$ and $\mathbb{Y} = (Y, d_Y)$ are two metric spaces. We define the correspondence $R \subset X \times Y$ with the following conditions:

- $\forall x \in X$, $\exists y \in Y$ *s.t.* $(x, y) \in R$
- $\forall y \in Y$, $\exists x \in X$ *s.t.* $(x, y) \in R$

Now the definition of the Gromov-Hausdorff distance is as follows:

$$d_{GH}(\mathbb{X}, \mathbb{Y}) = \frac{1}{2} \inf_{R \subset C(X,Y)} \max_{(x,y),(x',y') \in R} |d_X(x, x') - d_Y(y, y')|.$$

There is another definition for the Gromov-Hausdorff distance which is as follows:

Definition 18. *(Gromov-Hausdorff Distance)* [Mem07]
Let f (and g) be a function that map set X (Y) from space $\mathbb{X}$ ($\mathbb{Y}$) to a new space $\mathbb{Z} = (Z, d_Z)$. The Gromov-Hausdorff distance is defined as follows:

$$d_{GH}(\mathbb{X}, \mathbb{Y}) = \inf_{\mathbb{Z},f,g} d_H^{\mathbb{Z}}(f(X), g(Y)).$$

where $d_H^{\mathbb{Z}}$ is the Hausdorff distance at space $\mathbb{Z}$.

The Gromov-Hausdorff distance is a distance that can be used for comparing trees, but there is a problem with the Gromov-Hausdorff distance, and the problem is shown in the following theorem.

Theorem 2. [K A+18; Sch17]
Unless $P = NP$, there is no polynomial time algorithm to approximate the Gromov-Hausdorff distance between two metric trees with a factor of better than 3.

Remark 1. [K A+18]
The best known approximation algorithm is with the ratio of $O(\sqrt{n})$. In other words, there is an algorithm that can compute a distance $\hat{d}$ s.t. for a constant r we have:

$$d_{GH}(\mathbb{X}, \mathbb{Y}) \leq \hat{d} \leq r\sqrt{n}\ d_{GH}(\mathbb{X}, \mathbb{Y}).$$

In the first paper, we illustrate another distance between trees and use that distance to approximate the Gromov-Hausdorff distance between trees.

Another distance that is defined on phylogeny trees is called the Clustering Information Distance.

Definition 19. *(Clustering Information Distance)* [Smi20; TNB24; NEB10; Mei07]
In a tree, each edge cuts the leaves into two sets. If one of the sets has less than two elements we say that the cut is trivial. Assume that $C_{i,j}^1 = A_1|B_1$ is the cut of the edge e_{ij} from the tree number 1 and similar definition for $C_{i',j'}^2$. By considering the cut $C_{i,j}^1$ the probability that a leaf l belongs to A_1 is $P(A_1) = \frac{n_{A_1}}{n}$ where n_{A_1} is the number of leaves in A_1 and n is the total number of leaves in tree 1.

Now w.l.o.g the entropy related to $C_{i,j}^1$ is as follows:

$$H(C_{i,j}^1) = -P(A_1)\log(P(A_1)) - P(B_1)\log(P(B_1))$$

and the mutual information between $C_{i,j}^1$ and $C_{i',j'}^2$ is as follows:

$$I_{CL}(C_{i,j}^1; C_{i',j'}^2) = J(A_1, A_2) + J(A_1, B_2) + J(B_1, A_2) + J(B_1, B_2)$$

which for nontrivial clustering $C^1_{i,j}$ *and* $C^2_{i',j'}$ *is called the* mutual clustering information*, where:*

$$J(A,B) = P(A,B)\log(\frac{P(A,B)}{P(A)P(B)}).$$

The variation of information between the clustering $C^1_{i,j}$ *and* $C^2_{i',j'}$ *is defined as follows:*

$$VI(C^1_{i,j}, C^2_{i',j'}) = H(C^1_{i,j}) + H(C^2_{i',j'}) - 2I(C^1_{i,j}, C^2_{i',j'}),$$

which is a metric (proof in [Mei07]).

As the clustering information distance between trees is obtained by using different cuts in the trees, we use this distance in the third paper for finding the distance between hierarchical clustering trees.

In the third paper, we analyzed the time series by using the information that we earned from Swedish companies traded during the five years. Therefore, we take a glance over the time series and the definitions that we have used in the papers. For more information, we refer readers to [Tsa05].

1.4 Time Series in Finance

In statistics, a series of random variables that have been sampled in the same time intervals is called a *time series*. In this section, we propose some definitions that are used in my papers.

1.4.1 Asset Returns

In finance, we use the returns instead of the prices, because of the reasons explained after the definition of returns.

Definition 20. *(Simple Return)*
Simple return is the relative change of the price itself, and it is calculated as follows:

$$R_t = \frac{P_t - P_{t-1}}{P_{t-1}},$$

where P_t *is the price at time* t*.*

Now, consider that you put 1 kr in a bank, and the bank has a different option to give you interest for your money. They say that they give you 10 percent yearly but only 1 time at the end of each year, or 10 percent in

a year but they pay every 3 months, i.e., four times each year, or they pay every month and so on. And *log return* is calculated if the bank pays you continuously, or in other words, the time distance between two payments goes to 0. In the above example, if the bank pays continuously, then the money will be 1.10517 kr after one year and in this case, we have the following:

$$P_t = P_{t-1}e^r,$$

where r is the interest rate.

Definition 21. *(Log Return)*
Log return is the logarithm of the simple change of the price, and it is calculated as follows:

$$R_t = \log(\frac{P_t}{P_{t-1}}),$$

where P_t is the price at time t.

Using returns instead of prices has some advantages. First of all, as you see, the returns do not have scale. Secondly, the returns have interesting statistical features: they follow a stationary process, while the prices themselves are not stationary.

Definition 22. *(Weakly Stationary)*
We can predict a time series if it is Weakly Stationary *which means the average and the variance of the data do not depend on time. Also, its autocorrelation function depends only on the lag between the observations.*

Intuitively, in a weakly stationary time series, the way that time series changes over time does not change.

There are some methods to predict future values of a times series and to do so, we need to assign a model to the observed data. There are some ways to choose a model for a time series.

Definition 23. *(Auto-Regressive (AR) Model)*
In this model, for a natural number p each observation depends on a constant α_0, the p previous observations, and a white noise process $\{\varepsilon_t\}$[2] series ε_t with mean 0 and variance σ_ε^2. Therefore, the returns r_t follow an AR(p) process, if:

$$r_t = \alpha_0 + \alpha_1 r_{t-1} + \alpha_2 r_{t-2} + ... + \alpha_p r_{t-p} + \varepsilon_t,$$

[2] A sequence $\{\varepsilon_t\}$, that consists of random variables, is called a white noise process if they are independent and identically distributed and their mean and variance are finite.

where $\{\alpha\}_1^p$ *are the parameters of the model.*

Therefore, $AR(1)$ is defined as follows:

$$r_t = \alpha_0 + \alpha_1 r_{t-1} + \varepsilon_t.$$

By the assumption of being weakly stationary, $AR(1)$ has the following properties:

$$E(r_t) = \frac{\alpha_0}{1 - \alpha_1}$$

and variance r_t is earned as follows:

$$\text{Var}(r_t) = \frac{\sigma_\varepsilon^2}{1 - \alpha_1^2},$$

which requires that $|\alpha_1| < 1$.

One can extend the definition of $AR(p)$ to $AR(\infty)$ and find the following series:

$$r_t = \alpha_0 + \alpha_1 r_{t-1} + \alpha_2 r_{t-2} + ... + \varepsilon_t.$$

However, the problem is that we have an infinite number of parameters which can be challenging. Therefore the idea of the MA model is to change this series to the following:

$$r_t = \alpha_0 - \beta r_{t-1} - \beta^2 r_{t-2} - \beta^3 r_{t-2} - ... + \varepsilon_t, \tag{1.1}$$

and, similarly,

$$r_{t-1} = \alpha_0 - \beta r_{t-2} - \beta^2 r_{t-3} - \beta^3 r_{t-4} - ... + \varepsilon_{t-1}. \tag{1.2}$$

Substituting (1.2) in (1.1) we have

$$r_t = \alpha_0(1 - \beta) + \varepsilon_t - \varepsilon_{t-1}\beta.$$

In general, we find the following definition:

Definition 24. *(Moving-Average (MA) Model)*
In this model, for a natural number q *each value of the time series depends on a constant* β_0 *and the* q *previous errors. So,* $MA(q)$ *is defined as follows:*

$$r_t = \beta_0 + \varepsilon_t - \beta_1 \varepsilon_{t-1} - \beta_2 \varepsilon_{t-2} - ... - \beta_q \varepsilon_{t-q}.$$

$MA(1)$ is defined as follows:

$$r_t = \beta_0 + \varepsilon_t - \beta_1 \varepsilon_{t-1},$$

and by the weakly stationary assumption of r_t, we have that $E(r_t) = \beta_0$, and $Var(r_t) = (1 + \beta^2)\sigma_\varepsilon^2$.

We also can define the Vector Auto-Regressive (VAR) model similarly, which leads to the consideration of multivariate time series.

$$\mathbf{r_t} = \mathbf{A_0} + \mathbf{A_1 r_1} + ... + \mathbf{A_p r_p} + \epsilon_\mathbf{t},$$

where $\mathbf{A_0}$ is a vector of dimension n, and $\mathbf{A_j}$s are n by n matrices, and $\{\epsilon_\mathbf{t}\}$ is a series of vector of dimension n with mean vector 0 and covariance matrix K. This model is used in the third paper to analyze the time series that we earn by using some methods.

2 Summary of the Papers

In this section, I illustrate the four papers that I have worked on during my Ph.D. study at Stockholm University. In the first paper, the aim was to compare trees and define the distance (or dissimilarity) between trees. Working on networks, constructing networks from the data, and working on some properties of the networks was the main purpose of the second paper. In the third paper, by using the distance between the trees that I considered in my first paper and also by using the network definition that I worked with in the second paper, I analyzed the relationship between companies in the Swedish financial market during the last five years. Finally, the fourth paper presents several statistical tests on the structure of a Gaussian graphical model. The exact distributions of several test statistics were derived under the null hypothesis.

2.1 Summary of the First Paper

The distance between two trees is the main subject of the first paper [TW22]. There are a few methods to calculate the similarity between trees. However, none of them gives us enough information about the relationship between trees as was discussed in the previous section. Distance between trees has applications in cluster analysis [Lu79], neuron classification [LF12], and phylogenetic trees [Smi20].

The interleaving distance is a distance defined for finding a similarity between merge trees. A merge tree is a finite tree with a function that is

defined on the tree [MBW13]. The function is a monotonically decreasing function from the root to the leaves. It means that the value of the function for any father of a node is larger than the value for the node. Also, there is an edge from the root of the tree to infinity. Intuitively, we can construct a merge tree from a given tree by hanging the tree from a node. We define a function f by choosing 0 for the root and for any point in the tree the value of the function is negative of the distance from the point to the root. The interleaving distance is defined only between merge trees. For defining the interleaving distance first, we need to define two ε-compatible maps as follows:

Definition 25. *(ε-compatible maps) [MBW13]*
Two maps $\alpha : |T_1^f| \to |T_2^g|$ and $\beta : |T_2^g| \to |T_1^f|$ are ε-compatible if and only if the following conditions are satisfied:

- $g(\alpha(u)) = f(u) + \varepsilon, \quad \beta o \alpha(u) = u^{2\varepsilon}$
- $f(\beta(v)) = g(v) + \varepsilon, \quad \alpha o \beta(v) = v^{2\varepsilon}$.

If there exists an ε that the above four conditions satisfy, we can say that T_1^f and T_2^g are ε-interleave.

And, the definition of the interleaving distance between two merge trees is as follows:

Definition 26. *(Interleaving Distance) [MBW13]*
For two given merge trees T_1^f and T_2^g, the interleaving distance between them is defined as the infimum over all the ε s.t. T_1^f and T_2^g ε-interleave. It means:

$$d_I(T_1^f, T_2^g) = \inf\{\varepsilon, \exists a \text{ pair of } \varepsilon-\text{compatible maps between } |T_1^f| \text{ and } |T_2^g|\}$$

In this paper, we redefine the interleaving distance between two merge trees by only one map instead of two maps. We call the map ε-good map and we prove that the interleaving distance between two merge trees is ε if and only if there exists an ε-good map from T_1^f to T_2^g.

For finding an ε-good map between two merge trees first, we need to solve the following decision problem.

Decision Problem. For a given ε, whether the interleaving distance between two given merge trees is equal to ε or not?

In this paper we present two polynomial time algorithms for solving the above decision problem for a group of trees (with some conditions on trees). Furthermore, we prove that ε can be chosen from a list and present a fixed parameter tractable algorithm for finding the interleaving distance between two given merge trees with some conditions.

2.2 Summary of the Second Paper

In this paper [TL22], we look at one characteristic of the networks, which is the clustering coefficient, and extend its definition to the *Relative Clustering Coefficient* introduced in the paper. The clustering coefficient is useful in analyzing brain networks [Mas+18], social networks [CH20], distinguishing between a cancer network and a normal network [AT16], and also in transportation system [DL19]. When we look at real networks (networks for the real data), we expect that they have a lot of triangles, or in other words, we say that the network is highly clustered. This means that in a real network two friends of a person meet in some way and they become friends of each other.

Two types of clustering coefficients have been defined on networks, *Local Clustering Coefficient* and *Global Clustering Coefficient*. In the local clustering coefficient for any node V_i we look at the numbers of the nodes that are connected to it and divide it by the number of edges such that V_i is one endpoint of that. The global clustering coefficient is defined on the graph and it is calculated by dividing the number of triangles by the number of paths of length three.

In this paper, we define the relative clustering coefficient as follows. First, in the network that we construct from the data, we add weight to all the edges between any two nodes. We set the weight for the edge $e_{i,j}$ as 1, if the probability of having an edge between the nodes V_i and V_j is 1, otherwise the weight is 0. Let $|\triangle_3^1|$ be the number of triangles in the weighted network that all the edges of the triangle have the weight 1. We define $|\triangle_2^1|$ as the number of triangles in the weighted network that the weight of one of the edges is 0. Then, the relative clustering coefficient is defined as follows:

$$C_R = \frac{3 \times |\triangle_3^1|}{3 \times |\triangle_3^1| + |\triangle_2^1|}.$$

In this paper, we also considered the model that was proposed by M. E. J. Newman in [New03]. The model is as follows:

Model 1. *[New03]*

- *There are n individuals in the network that are divided into m different groups.*
- *Each individual belongs to at least one group.*
- *Individuals belong to groups randomly.*
- *If two individuals belong to a group, with the probability of p, they are connected. And with the probability of* $1 - p$ *they are not connected.*

Using this model, we can ascertain the importance of the relative clustering coefficient C_R over the global clustering coefficient C. Namely, if we look at the number of triangles in the network which is constructed by using this model, i.e., the denominator in the definition of the global clustering coefficient, we consider all the triangles even those which connect two individuals from different groups, however, the probability of having an edge between two individuals from different groups is zero. On the other side, the application of the relative clustering coefficient solves the problem. Moreover, we proved that when n goes to infinity, C_R goes toward p.

2.3 Summary of the Third Paper

In this paper [TNB24], we use the distance defined between trees in the first project and the conversion of the data into networks in the second project, we analyze the relationship between the companies traded on the Swedish stock exchange during the past five years. Here we use two methods for finding the relationship between the financial returns of the companies and constructed networks between companies based on them. One of the methods is the *Pearson Correlation Coefficients (PCC)*, which is used to compute the Pearson Correlation Coefficients Dissimilarity (PCCD) matrix that indicates the dissimilarity between companies. Another method is the *Generalized Variance Decomposition (GVD)* method which was used to find the Generalized Variance Decomposition Dissimilarity (GVDD) matrix that indicates the dissimilarity between companies. And, we want to see if there are any changes in the time series obtained by different methods during the past five years, especially during the COVID-19 pandemic.

In this paper, first, we choose a window of three months which consists of 63 trading days, and compute the log returns of the adjusted closing prices for 28 top companies traded on the Swedish capital market. Then, we consider the data of returns separately at several windows, which are obtained by shifting the window by one day to construct the next window. At each time t, we have a 63×28 matrix of stock returns, which are used to construct two distance matrices by applying the PCCD and GVDD methods as follows:

PCCD Method: [Man99]
First, we construct the ρ matrix such that each element of the matrix is

calculated by:

$$\rho_{i,j}^{t} = \frac{cov(C_i^t, C_j^t)}{var(C_i^t) \times var(C_j^t)}.$$

where C_i^t indicates the i-th column of the return matrix at time t and the distance matrix at time t is given by:

$$h_{PCCD}^{t,ij} = \sqrt{2(1 - \rho_{i,j}^{t})}.$$

GVDD Method: [DY14]
By using the GVDD method we consider the following H-step-ahead generalized variance decomposition matrix (GVD) at time t:

$$\nu_{ij}^{t} = \frac{\kappa_{jj}^{-1} \sum_{s=0}^{H-1} \left(\mathbf{e}_i^{'} \mathbf{\Sigma}_s \mathbf{K} \mathbf{e}_j \right)^2}{\sum_{s=0}^{H-1} \left(\mathbf{e}_i^{'} \mathbf{\Sigma}_s \mathbf{K} \mathbf{\Sigma}_s^{'} \mathbf{e}_i \right)}$$

where $\mathbf{K}$ is the covariance matrix of the error process and $\mathbf{\Sigma}_s$ is the coefficient of the infinitive moving average (MA) representation of the VAR(p) model, $\mathbf{e}_i$ is a vector of length n with 1 at the ith element and 0 elsewhere, and κ_{jj} is the jjth element of matrix K. In this project, we consider p = 1. By using the following matrix,

$$\bar{\nu}_{ij}^{t} = \frac{\nu_{ij}^{t}}{\sum_{j=1}^{n} \nu_{ij}^{t}}$$

we find the GVDD distance matrix, as follows:

$$h_{GVDD}^{t,ij} = \sqrt{2(1 - \bar{\nu}_{i,j}^{t})}$$

Using these two matrices, $h_{GVDD}^{t,ij}$ and $h_{PCCD}^{t,ij}$, we construct two networks between the 28 companies in Sweden at each window. Then, each network's center is found using the data from the past five years. Interestingly, both methods determine *Investor AB* as the most influential company among 28 top companies traded on the Swedish stock exchange.

Furthermore, we find the hierarchical clustering of the data at each window by using two methods and calculate the distance between hierarchical clustering trees during the past five years. We observe that by using the

GVDD method the distance between hierarchical clustering is higher than when the PCCD method is applied. Also, by fitting a $VAR(2)$ model (the order 2 is determined by using Hannan-Quinn information criterion in [HQ79]) to this vector time series we observe that the values of both the PCCD method and the GVDD method have a positive correlation to their previous values. However, both variables have no significant impact on the OMX index, the index of the Swedish capital market. Also, we find that the changes in the structure of trees happen mostly during the COVID-19 pandemic.

2.4 Summary of the Fourth Paper

In this paper [BT23], we consider a Gaussian random vector $X = (x_1, x_2, ..., x_p)$ with the multivariate Gaussian distribution $\mathcal{N}(\boldsymbol{\mu}, \mathbf{K})$ where $\mathbf{K}_{i,j} = cov(X_i, X_j)$ with density function given by [BDP19]

$$f(X) = (2\pi)^{-\frac{p}{2}} |\mathbf{K}|^{-1} \exp\left(-\frac{1}{2} (X - \boldsymbol{\mu})^\top \mathbf{K}^{-1} (X - \boldsymbol{\mu})\right).$$

A *Gaussian Graphical Model (GGM)* is defined by using the structure of the inverse covariance matrix $\mathbf{K}^{-1}$ as follows.

Definition 27. *(Gaussian Graphical Model) [Kal+19]*
We say that a random vector $X = (x_1, x_2, ..., x_p)$ follows a GGM with the graph $G = (V, E)$ with the set of vertexes $V = \{v_1, ..., v_p\}$ associate with the nodes X and $E \subset \{\{v_i, v_j\},\ where\ \ v_i, v_j \in V\}$ where

$$\{v_i, v_j\} \notin E \quad \textit{iff} \quad w_{i,j} = 0,$$

with $w_{i,j}$ the (i,j) elements of the inverse covariance matrix $\mathbf{K}^{-1}$, also known as the precision matrix.

Using the relation between a Gaussian graph and an inverse covariance matrix, we derive statistical tests on specific graphical models. Exact test theory is derived for the following GGMs:

1. The precision matrix is the matrix corresponding to the AR(1) model.

2. The precision matrix is block-diagonal. In this case, a graph consists of several connected components.

3. The precision matrix is related to the star graph.

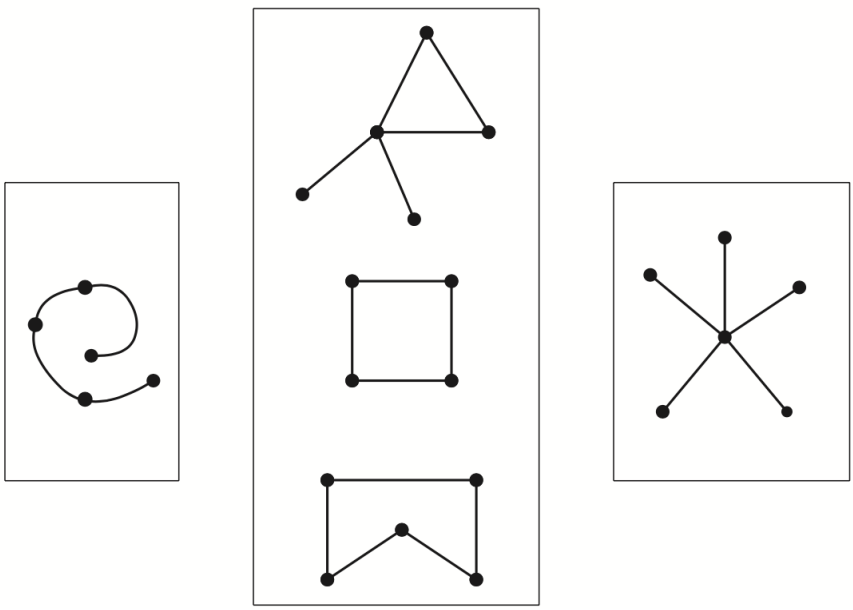

Figure 3: Three types of graphs: left-hand-side plot corresponds to the AR(1) structure of the precision matrix, middle plot the block-diagonal structure of the precision matrix, right-hand-side plot factor structure. From "Exact Test Theory in Gaussian Graphical Models", by Olha Bodnar, Elena Farahbakhsh Touli, 2023, Journal of Multivariate Analysis 196: 105185. Copyright 2024 by Elena Farahbakhsh Touli with permission.

Figure 3 depicts examples of graphs that correspond to several types of the precision matrix considered in the paper. All of the three types of graphs are widely applied in different fields of science.

Using the data $\mathcal{X} = \{X_1, ..., X_n\}$ with $n > p$, which consists of independent and identically distributed random vectors, we derive exact test theory for each of the three null hypotheses considered in the paper. The exact distribution of the test statistic derived for each of the considered types of a Gaussian graph is obtained under the null hypothesis. We compare the new suggested approaches with the existent benchmark method, which is based on testing each component of the precision matrix separately and makes use of the Fisher transformation. In contrast to the benchmark approach, which is based on asymptotic tests, the new methods use the exact distribution. As a result, they control the probability of the overall type I error rate at the desired level, while the application of the asymptotic benchmark approach usually leads to higher values of the realized probability of the type I error rate. Finally, within an extensive simulation study, we show that the new tests are more powerful in detecting deviations from the null hypothesis in comparison to the asymptotic approach.

3 Conclusion and Further Work

This thesis is comprised of four key papers. The first paper redefines the interleaving distance between merge trees, providing a novel framework for measuring similarity. The second paper introduces the concept of the relative clustering coefficient, expanding upon traditional clustering metrics. In the third paper, we analyze the distance between hierarchical clustering trees obtained from different time windows over the past five years, using single linkage clustering exclusively. Future work could explore comparisons between single linkage and average linkage clustering methods. The final paper focuses on Gaussian graphical models, employing three types of graphs for the structure of the precision matrix. Further research may investigate additional types of precision matrices to enhance the model's robustness.

4 Svensk Sammanfattning

Att arbeta med olika problem relaterade till grafteori kombinerat med statistik är huvudsyftet med denna avhandling. I papper I arbetade vi med avståndet mellan träd och definierade en annan definition för Interleaving-avståndet som redan introducerats för att bestämma avståndet mellan sammanslagna träd. Den nya definitionen baserades på endast en avbildning från ett av träden till ett annat. På så vis kunde vi få algoritmer som är hanterbara med fasta parametrar för att hitta Interleaving-avståndet mellan sammanslagna träd under vissa förutsättningar.

I papper II arbetade vi med klusterkoefficienten på nätverk. Klusterkoefficienten indikerar hur benägna nätverkets noder är att bilda en triangel. Vi introducerade en annan klusterkoefficient som vi kallade för relativ klusterkoefficient. Till sist föreslogs vikten av den relativa klusterkoefficienten genom en model.

I papper III arbetade vi med finansiella relationer mellan företag i Sverige och använde två metoder (Pearsons korrelationskoefficient (PCC) och generaliserad variansdekomposition (GVD)). Vi tillämpade sedan dessa metoder på finansiella data bestående av dagliga avkastningar på de 28 aktierna som ingår i beräkningen av OMX-indexet (index för den svenska kapitalmarknaden).

Gaussisk grafisk modell är huvudämnet i artikel IV. I denna artikel överväger vi tre typer av precisionsmatriser och utvecklar för varje typ av precisionsmatris den exakta testteorin. Slutligen jämförs de nya metoderna med referensmetoden genom en omfattande simuleringsstudie.

Bibliography

[AT16] H. R. Arabnia and Q. N. Tran. *Emerging Trends in Applications and Infrastructures for Computational Biology, Bioinformatics, and Systems Biology: Systems and Applications*. 1st. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2016. ISBN: 0128042036.

[Aro+92] S. Arora et al. "Proof Verification and Hardness of Approximation Problems". In: *FOCS*. IEEE Computer Society, 1992, pp. 14–23. ISBN: 0-8186-2900-2. URL: http://dblp.uni-trier.de/db/conf/focs/focs92.html#AroraLMSS92.

[Bil05] P. Bille. "A survey on tree edit distance and related problems". *Theoretical Computer Science* 337.1 (2005), pp. 217–239.

[BT23] O. Bodnar and E. F. Touli. "Exact test theory in Gaussian graphical models". *Journal of Multivariate Analysis* 196 (2023), p. 105185. ISSN: 0047-259X. DOI: https://doi.org/10.1016/j.jmva.2023.105185. URL: https://www.sciencedirect.com/science/article/pii/S0047259X23000313.

[BDP19] T. Bodnar, H. Dette, and N. Parolya. "Testing for Independence of Large Dimensional Vectors". *The Annals of Statistics* 47 (Aug. 2019), pp. 2977–3008. DOI: 10.1214/18-AOS1771.

[CH20] C.-Y. Chen and J.-J. Huang. "A Novel Centrality for Finding Key Persons in a Social Network by the Bi-Directional Influence Map". *Symmetry* 12.10 (2020). ISSN: 2073-8994. DOI: 10.3390/sym12101747. URL: https://www.mdpi.com/2073-8994/12/10/1747.

[Cor+01] T. H. Cormen et al. *Introduction to Algorithms*. 2nd. The MIT Press, 2001. ISBN: 0262032937. URL: http://www.amazon.com/Introduction-Algorithms-Thomas-H-Cormen/dp/0262032937%3FSubscriptionId%3D13CT5CVB80YFWJEPWS02%26tag%3Dws%26linkCode%3Dxm2%26camp%3D2025%26creative%3D165953%26creativeASIN%3D0262032937.

[DY14] F. X. Diebold and K. Yılmaz. “On the network topology of variance decompositions: Measuring the connectedness of financial firms”. *Journal of Econometrics* 182.1 (2014), pp. 119–134.

[DL19] Z. Doukha’ and I. Loumachi. “Chapter 15 - Secure Data Dissemination for Smart Transportation Systems”. In: *Smart Cities Cybersecurity and Privacy*. Ed. by D. B. Rawat and K. Z. Ghafoor. Elsevier, 2019, pp. 217–225. ISBN: 978-0-12-815032-0. DOI: https://doi.org/10.1016/B978-0-12-815032-0.00015-9. URL: https://www.sciencedirect.com/science/article/pii/B9780128150320000159.

[ES97] U. M. E. Belogay C. Cabrelli and R. Shonkwiler. “Calculating the Hausdorff distance between curves”. *Information Processing Letters* 64 (1997), pp. 17–22.

[Gal13] M. Gallinger. *Gröbner Bases: Ideal Membership and Graph Colouring*. Lakehead University, 2013.

[GJ79] M. R. Garey and D. S. Johnson. *Computers and Intractability: A Guide to the Theory of NP-Completeness (Series of Books in the Mathematical Sciences)*. First Edition. W. H. Freeman, 1979. ISBN: 0716710455. URL: http://www.amazon.com/Computers-Intractability-NP-Completeness-Mathematical-Sciences/dp/0716710455.

[HQ79] E. J. Hannan and B. G. Quinn. “The determination of the order of an autoregression”. *Journal of the Royal Statistical Society: Series B (Methodological)* 41.2 (1979), pp. 190–195.

[JWZ95] T. Jiang, L. Wang, and K. Zhang. “Alignment of Trees - An Alternative to Tree Edit”. *Theor. Comput. Sci.* 143.1 (1995), pp. 137–148.

[K A+18] P. K. Agarwal et al. “Computing the Gromov-Hausdorff Distance for Metric Trees”. *ACM Trans. Algorithms* 14.2 (2018), 24:1–24:20.

[Kal+19] V. A. Kalyagin et al. “Loss function, unbiasedness, and optimality of Gaussian graphical model selection”. *Journal of Statistical Planning and Inference* 201 (2019), pp. 32–39.

[LF12] R. Lefort and F. Fleuret. “A tree-based distance between distributions: Application to classification of neurons”. In: *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2012, pp. 2237–2240. DOI: 10.1109/ICASSP.2012.6288358.

[Lu79] S.-Y. Lu. "A Tree-to-Tree Distance and Its Application to Cluster Analysis". *IEEE Transactions on Pattern Analysis and Machine Intelligence* PAMI-1.2 (1979), pp. 219–224. DOI: 10.1109/TPAMI.1979.6786615.

[Man99] R. N. Mantegna. "Hierarchical structure in financial markets". *The European Physical Journal B-Condensed Matter and Complex Systems* 11.1 (1999), pp. 193–197.

[Mas+18] N. Masuda et al. "Clustering coefficients for correlation networks". English. *Frontiers in Neuroinformatics* 12 (Mar. 2018). Publisher Copyright: © 2018 Masuda, Sakaki, Ezaki and Watanabe. ISSN: 1662-5196. DOI: 10.3389/fninf.2018.00007.

[Mei07] M. Meilă. "Comparing clusterings – an information based distance". *Journal of Multivariate Analysis* 98.5 (2007), pp. 873–895. ISSN: 0047-259X. DOI: https://doi.org/10.1016/j.jmva.2006.11.013. URL: https://www.sciencedirect.com/science/article/pii/S0047259X06002016.

[Mem07] F. Memoli. "On the use of Gromov-Hausdorff Distances for Shape Comparison". In: The Eurographics Association, 2007, pp. 81–90.

[MBW13] D. Morozov, K. Beketayev, and G. H. Weber. "Interleaving Distance between Merge Trees". In: *Workshop on Topological Methods in Data Analysis and Visualization: Theory, Algorithms and Applications*. 2013.

[New03] M. E. J. Newman. "Properties of highly clustered networks". *Phys. Rev. E* 68 (2 Aug. 2003), p. 026121. DOI: 10.1103/PhysRevE.68.026121. URL: https://link.aps.org/doi/10.1103/PhysRevE.68.026121.

[NEB10] X. V. Nguyen, J. Epps, and J. Bailey. "Information Theoretic Measures for Clusterings Comparison: Variants, Properties, Normalization and Correction for Chance." *J. Mach. Learn. Res.* 11 (2010), pp. 2837–2854. URL: http://dblp.uni-trier.de/db/journals/jmlr/jmlr11.html#NguyenEB10.

[Sch17] F. Schmiedl. "Computational Aspects of the Gromov–Hausdorff Distance and its Application in Non-rigid Shape Matching". *Discrete & Computational Geometry* 57.4 (2017), pp. 854–880.

[Smi20] M. R. Smith. "Information theoretic generalized Robinson–Foulds metrics for comparing phylogenetic trees". *Bioinformatics* 36.20 (2020), pp. 5007–5013. ISSN: 1367-4803.

[TW22] E. F. Touli and Y. Wang. "FPT-Algorithms for computing Gromov-Hausdorff and interleaving distances between trees". *Journal of Computational Geometry* 13.1 (2022), pp. 89–124.

[TL22] E. F. Touli and O. Lindberg. "Relative Clustering Coefficient". *Journal of Algorithms and Computation* 54.1 (2022), pp. 99–108. ISSN: 2476-2776. DOI: 10.22059/jac.2022.88373.

[TNB24] E. F. Touli, H. Nguyen, and O. Bodnar. "Monitoring the Dynamic Networks of Stock Returns with an Application to the Swedish Stock Market". *Computational Economics* (2024), pp. 1–18. ISSN: 0047-259X. DOI: https://doi.org/10.1007/s10614-024-10616-2. URL: https://link.springer.com/article/10.1007/s10614-024-10616-2.

[Tsa05] R. Tsay. *Analysis of financial time series*. 2. ed. Wiley series in probability and statistics. Hoboken, NJ: Wiley-Interscience, 2005. XXI, 605. ISBN: 978-0-471-69074-0. URL: http://gso.gbv.de/DB=2.1/CMD?ACT=SRCHA&SRT=YOP&IKT=1016&TRM=ppn+483463442&sourceid=fbw_bibsonomy.

[WF94] S. Wasserman and K. Faust. *Social Network Analysis: Methods and Applications*. Vol. 8. Cambridge university press, 1994. URL: http://scholar.google.com/scholar.bib?q=info:gET6m8icitMJ:scholar.google.com/&output=citation&hl=en&as_sdt=0,5&as_vis=1&ct=citation&cd=0.

[ZJ94a] K. Zhang and T. Jiang. "Some MAX SNP-hard results concerning unordered labeled trees". *Information Processing Letters* 49.5 (1994), pp. 249–254. ISSN: 0020-0190. DOI: https://doi.org/10.1016/0020-0190(94)90062-0. URL: https://www.sciencedirect.com/science/article/pii/0020019094900620.

[ZJ94b] K. Zhang and T. Jiang. "Some MAX SNP-hard results concerning unordered labeled trees". *Information Processing Letters* 49.5 (1994), pp. 249–254.