UNIVERSITÀ DEGLI STUDI DI MILANO

FACOLTÀ DI SCIENZE POLITICHE, ECONOMICHE E SOCIALI

Master's Degree Course in Data Science for Economics

EUROPEAN BANKS SYSTEMIC RISK BETWEEN 2021 AND 2023: AN ANALYSIS THROUGH BAYESIAN NETWORKS

Supervisor: Prof Luca ROSSINI

Co-supervisor: Prof Silvia SALINI,

Thesis by:
Nicolò PIGNATELLI
Student ID: 13831A

Academic Year 2023-2024

"Economists set themselves too easy, too useless a task if in tempestuous seasons they can only tell us that when the storm is long past the ocean is flat again."
— John Maynard Keynes

Acknowledgments

I would like to thank Prof. Rossini and Prof. Salini for valuable and timely advice.

Contents

List of Figures

Abstract

The recent crises that hit the world economy have emphasized the need to develop models capable of predicting these events and their effects on all actors in the system. A particularly important sector in the outbreak of a crisis is the banking sector, from which financial instability often originates and grows, then spreads to the entire economic system through a domino effect. This thesis analyses the systemic risk of the European banking sector between 2021 and 2023, through CDS spreads. To estimate the evolution of the risk during the two-year period, Bayesian Networks were implemented, thanks to which the probabilities that a bank will go through a period of stress conditioned by the stress of another bank were inferred. The results show a sharp increase in February 2022, when Russia started the invasion of Ukraine. After about four or five months, the level returned to the one before the outbreak of the crisis. According to this analysis, the banks most responsible for the propagation of the stress were *BNP Paribas* and *Danske Bank A S*. Moreover different robustness checks with rolling window or splitted data were provided.

Chapter 1

Introduction

1.1 Why and how to analyze the systemic risk of banks

In recent years a lot of periods of instability for the world economy happened: in 2007-2008 a global financial crisis, in 2010-2011 the European sovereign debt crisis, in 2020 the effects of the COVID-19 pandemic and in 2022 Russia invaded Ukraine. These crises could have been different in the causes and in the consequences, but still hit harsh all the actors in the market. One way to try to counter the losses induced by crises is to develop models and tools ideally capable of forecasting the effects of the resulting instability, as it would allow all the entities to prepare adequately if the worst scenario happens.

One interesting point of view from which to study financial crises is the banking sector. In particular during the first two crises mentioned above, the instability originated and grew in the financial sector and spread in the real economy from banks (4). In fact, it is being studied how the failure of financial institutions has a snowball effect on the non-financial world through restrictions on access to liquidity and worsening of accounts and therefore authors tend to analyze the interaction between the two parties (5; 16). It is thus interesting to study the so called *systemic risk* of the banking sector, as it is one fundamental ingredient of a financial crisis. Systemic risk is defined as the risk of distress in a large proportion of entities, hence, reported to the case study in this analysis, studying systemic risk means considering an event during which there is a loss in a large proportion of banks (16). It is thus not the same as systematic risk, since the latter refers to the risk that generally comes from the system, with no reference to the magnitude and the prevalence, and it is referred to a single asset or entity. On the contrary, the notion of systemic risk is inherently tied with the concept of dependency among the entities considered, as the probability of many banks being in distress is clearly affected by how they are connected to each other. If there are strongly interconnected, it is evident that the instability of one has a cascading effect on the others.

Given the definition of systemic risk, it would be appropriate to model all the interactions among the entities in the system, since it would enable to study thoroughly how any change in the financial

health of a bank alters the possibility of distress of another bank. It poses a challenge since it is difficult to work with different entities, and in fact there are different methods for studying the dependence structure when the variables considered are a few with respect to the ones for studying it when the set is large (16). To solve this issue, one can simplify the problem by considering the individual firm and aggregating all the others in a general market index. Some famous measures of systemic risk resulting from these methods are the Conditional Value at Risk (CoVaR) and the Marginal Expected Shortfall (MES). They have the advantage of being parsimonious and simple, but may lose important information by aggregating all the other firms in a single index. The idea is thus to overcome these methods and to keep all the information by considering all the entities separately. The difficulty of managing many entities however rises even more considering that the dependence patterns in financial time series are usually nonlinear and asymmetric, tend to change over time, structural breaks are all in all frequent and any normality hypothesis is usually discarded (4). It is therefore comprehensible why the commonly used correlation coefficient fails to capture these specificities.

In the literature there are two statistical tools that are used to deal with this complex scenario: the copula methodology (4; 16) and *Bayesian Networks* (18). This analysis was carried out with the latter, whose definition, properties, and advantages are described later (chapter 3), but briefly a Bayesian Network is a model that belongs to the family of Probabilistic Graphic Models that can handle robustly the joint probability distribution of a large number of variables and is excellent for analyzing dependency relationships between variables (11).

Once defined the statistical method used to tackle this issue, a crucial matter is deciding which data are suitable as an input. A popular answer is to use a financial asset named *Credit Default Swap*, shortly CDS, which detailed description is postponed later (chapter 1.2). Briefly a CDS is a contract between two parties, one that buys protection and one that sells protection. Typically, the former has a claim on a third party, called the reference entity, and wants to be insured against its possible default. The seller is willing to provide it in return for a periodic payment and therefore to bear the risk and any loss in the unfortunate event. The cost of the insurance is the CDS spread, and this analysis is in fact a time series analysis of CDS spreads of European banks (5). These assets are thus very useful for studying systemic risk because they are inherently tied with the market perception of risk of default of the reference entity. In fact, under some simplifying assumptions, CDS spread in basis points (the usual unit of measure) for the reference entity i at time t can be derived as (16):

$$S_{i,t} = 100^2 P_{i,t}^{\mathrm{Q}} L_{i,t}, \tag{1}$$

where $S_{i,t}$ is the spread in basis points, $P_{i,t}^{\mathrm{Q}}$ is the implied probability of default and $L_{i,t}$ is the loss given default.

The same expression can also be written in terms of objective probability of default, $P^{\mathrm{P}}_{i,t}$:

$$S_{i,t} = 100^2 P^{\mathrm{P}}_{i,t} M_{i,t} L_{i,t}, \tag{2}$$

where $M_{i,t}$ is the market price of risk. It is therefore evident that an increase in the spread can be explained by a rise in the objective probability of default, in the market price of risk or in the loss given default, and any of these three scenarios signals a worsening situation for the reference entity. This does not exhaust the reasons for using CDS to study systemic risk. First of all, it is a widely used asset. Recent estimation suggest amounts of contracts in circulation well in excess of one trillion dollars every quarter, even if historically it used to be much more (13). They are in fact the most used derivative for credit risk hedging, i.e. for mitigating the risk derived by lending money to an entity that may default or undergo any circumstance that makes it impossible to meet payment deadlines. This makes CDS the ideal tools for analyzing insiders' perceptions of the credit market as a whole and the situation of an individual company. Another key advantage of CDS is that they incorporate credit risk information faster than bonds (15). To be fair, it should be added that it is usually not possible to find reliable and well-structured data for free and in fact this analysis is based on information that has been purchased from a data provider.
To sum up, the various recent financial crises have stimulated the study of models capable of predicting the economic consequences of these events. One area that has proven to be crucial for understanding how these crises work is banking, and it is therefore useful to study the associated systemic risk. The method chosen to try to avoid oversimplification and loss of information is the estimation of Bayesian Networks that have as input data the CDS spreads of the banks as reference entities. In the remainder of this introduction, firstly it is described the functioning of CDS and it is illustrated the reasons why and how the use of typical complexity theory tools, in particular networks, can be helpful in the study of financial matters. Following this, the second chapter deals with data description, preprocessing, analysis and the transformations required to make them suitable for estimating Bayesian Networks. The third chapter describes in detail what is a Bayesian Network, the advantages of this model and how it can be estimated. In the fourth chapter, the results obtained are illustrated and finally, in the fifth chapter, conclusions are drawn.

1.2 What are CDS and how they work

A Credit Default Swap, shortly CDS, is an OTC derivative that allows to perform credit risk management (13). OTC means Over-the-Counter, i.e. a trade negotiated and concluded between two parties, whereas a derivative is a contract that derives its value from the performance of an underlying entity. Credit risk is the risk that originates from the lending of money, as the debtor may default or generally do not respect the deadlines of the agreed payments. CDS are the most popular and liquid assets for credit risk management, so the associated market is paid attention to by the various stakeholders to understand the situation in the credit market as a whole. The two

parties involved in a usual CDS negotiation are known as Protection Buyer and Protection Seller. Typically, the former lent money to a third party and wants to be insured against loss due to the debtor's inability to fulfill its obligations. In short, the buyer wants to transfer this credit risk and the protection seller is the one who is willing to bear it in return for a monetary compensation. The debtor of the buyer is known as the reference entity. This is how a CDS originates, and so, as already stated, a usual reason for a buyer to stipulate this contract is hedging, since it holds the underlying obligation (i.e. the credit instrument). Anyway, once created, the CDS could also be exploited for speculating on the current evaluation of the credit risk. There exist single-name CDS, the ones considered in this discussion, which refer to only one entity, and CDS indices, which could have multiple underlying entities. Generally, reference entities are not involved under any circumstances in any negotiation related to their debt.

The conclusion of a CDS between the two parties thus stems primarily from the protection buyer's desire to mitigate the credit risk associated with a loan to a certain entity. This risk is quantified and the resulting amount is known as the notional amount. Once a Protection Seller is found, the Buyer usually pays immediately an upfront amount and then a fixed one on that notional every quarter. The premium established during a trade is composed by two parts: the upfront amount, a payment executed immediately from the Buyer as the negotiation is concluded and is proportional to the risk that the Seller bears, and the standardized fixed coupon, that is paid regularly every quarter during all the lifetime of the contract and is usually set at 100 or 500 basis points (shortly bps). Basis points are the usual unit of measure in the credit derivative markets instead of percentages (100 bps are equal to 1%).

The upfront amount with the fixed coupon form the par CDS spread, i.e. the spread that would set the present value of the trade equal to 0. Since the coupon is paid throughout all the existence of the CDS and the value of money changes over time, it is necessary to discount the future cash flow to compute the actual value of the premium. The par spread reflects the market perceptions on the credit risk of the reference entity: if it shows signs of difficulty in paying its debts, the spread will rise, as a possible protection seller will demand a higher premium for a higher perceived risk, and vice versa. Given that the coupon is fixed however, the upfront amount is the one used to set the present value equal to the par. The spread also depends on the Tenor of the contract, that is the duration. Usually, the most liquid and frequently quoted part of the credit curve, the one that links the duration with the spread, are the 5 year terms. This curve is constructed to determine the price of a CDS and hence the underlying risk. Another fundamental notion of a CDS is the credit event that, if triggered, forces the Seller to cover any loss related to this circumstance and the contract to cease to exist. There are specific events that are taken into account as possible triggers, for example bankruptcy of the reference entity, failure to pay, restructuring and others.

It is now important to deepen the topic of how CDS are priced. First of all, CDS contracts are divided in two main components. The first component is the premium leg, the part during which the Buyer pays the premium regularly. This component may end with the natural expiration of the contract or because a credit event is triggered, when the Seller covers the loss. This is known as the

contingent leg. This clarification is very important for CDS pricing since to compute the present value of this derivative it is not only necessary to discount future cash flows, but also to take into account that they may not take place because the credit event is triggered and the Seller has to pay the Buyer. It is therefore necessary to assess the probability of default of the reference entity and the loss given default.

Before 2009, CDS were quoted and traded in par spreads, whose definition can be updated to that spread that makes equal the present value of the premium leg and the present value of the contingent leg. As the par spread contains all the information about the health of the reference entity, it was used also in practice to conclude the trade. Before 2009, in fact, if a Buyer accepted an offer of a certain amount of bps to receive protection, that quantity would had also been the regular premium to pay every quarter. This simple mechanism, however, was changed after it contributed to the outbreak of the 2008 crisis and this is also the reason why the notional outstanding of CDS has dramatically decreased after having reached the $60 trillion mark before the subprime crisis (5). After being created in the mid-1990s, prompted in particular by banks, which had an interest in insuring the risks associated with their lending activities, CDS became a complex and opaque instrument. In particular they were used to concentrate risk and this led the competent authorities in 2009 to intervene and standardize the market.

On 8 April the Big Bang Protocol came into force, such an important event in the history of these derivatives that it must be taken into account when analyzing time series that include that date (16). The most important feature introduced were the standardized coupon rates and the quoting conventions. At this point, the exchange took place with fixed coupons and no longer with par spreads. It therefore became necessary to add the upfront amount already mentioned to adjust the value to the actual underlying risk. In this way, for example, if the credit risk is greater than the one indicated by the standardized coupon, the Protection Seller will demand an upfront payment, otherwise the Buyer will receive it. The upfront can be approximately derived as:

$$UF \approx (FixedCoupon - ParSpread) \times RPV01, \tag{3}$$

where RPV01 is the sum of the discount factor weighted by survival probabilities, i.e. the probability that the reference entity does not default. It is used to integrate the discount factor and the credit risk over time. From 2009 onwards, CDS are usually quoted in *Conventional Spread*, which today are the ones that take into account the upfront amount and the regular premium. From a mathematical point of view, they can be inserted in Equation (3) instead of Par Spread, but they are computed with a different RPV01 that is constrained by certain precise assumptions about credit and interest rate yield curves. This is the reason why this analysis was carried out with conventional and not par spread.

Understanding how CDS work is crucial to justify why it is considered an excellent tool for studying systemic risk and for assessing the goodness of the choices made during this analysis, in particular which data to work with. The conventional spreads, after an initial processing, will be the data

on which the algorithms derive the final dataset to be provided as input to the Bayesian Networks. After presenting the data on which this work will be based, the next section illustrates the usefulness of some mathematical tools in the study of financial issues, in particular networks as typical means for dealing with complex systems.

1.3 Applying networks for studying finance

A complex system is usually understood as a system for which it is difficult to derive the collective behavior from a knowledge of its components (2). The definition itself explains that the analysis of these objects poses challenges. The science that studies the theory of complex systems is called complexity theory. The tool that makes it possible to model the components of any complex system and the interactions between them is a network. The study of any complex system therefore has as a fundamental step the definition of the underlying network, representable by a mathematical object named graph, composed of nodes, that are the components of the systems, and edges, that represent the links between the nodes. What has made network science, the specific discipline that studies networks, such a thriving field of study recently, apart from the technological development that has made the use of huge masses of data possible, has also been the discovery that the principles and laws that govern the networks of these phenomena, which are so varied and heterogeneous in the real world, are in the final analysis the same.
The financial system is also suitable as an object of study for network science. When represented with a graph, the financial institutions are the nodes and the edges can be modeled according to the specific aim of the study. A network is in fact an excellent tool for analyzing cascading failures, i.e. those events during which, following the failure of one component, there is a domino effect whereby other components follow the same fate. Cascading failures are typical of complex systems and are precisely those events that occur at the beginning of a major financial crisis. To try to predict these events, it is useful to know the structure of the system in which one fears they might occur. Furthermore, it is useful to understand how these components interact. By modeling the structure of the network and the interconnections between nodes, it is possible to assess the robustness of the system. This is exactly what one wants to know when trying to understand how high the systemic risk of the financial system is. What defines a network is therefore basically a set of interconnected entities. Interconnectedness can inevitably cause vulnerability, as what happens to one node has repercussions on others (non-locality). This is precisely what can be observed during a major financial crisis, for example in 2008 the crisis burst in the USA and rapidly spread all over the world.
Networks thus seem to be a useful means for studying systemic risk and, in general, certain phenomena typical of financial crises. In [3], an attempt is made to elucidate the situation in academia regarding the use of complexity theory and the economic-financial world and why it should be encouraged. According to the authors, the initial impetus for exploring this topic originated in the complete failure of traditional economic theory to predict the 2008 crisis and its disruptive consequences. This has led some scholars to delve into some of the typical tools of complexity theory

to try to improve their understanding of the mechanisms that give rise to financial crises to be better prepared for the following scenarios. According to the authors, it is impossible to predict the individual events that initiate the domino effect typical of financial crises, and indeed this cannot be the goal to pursue. What usually happens is that, before the outbreak of the crisis, certain endogenous factors make the system less resilient and eventually favor the cascade effect. It would be therefore possible to analyze this gradual and hidden loss of resilience to anticipate the causes and consequences of a crisis.

In order to detect these changes, empirical indicators can be used to measure the resilience of the system and try to understand when the tipping point will occur, i.e. the single event that triggers the avalanche effect. These indicators include, for example, the correlation between nodes in a network. This idea has already been taken up by some insiders and in fact the use of complexity theory tools in the economic-financial field has already been adopted by non-academic institutions. For example, with regard to the area covered by this analysis, the Bank of International Settlements has developed a framework to measure the systemic risk associated with each bank and classify the riskiest ones (G-SIBs) based on the interconnection between banks. This classification is a source of study and comparison even in the academic field when it comes to the associated topic (15). Furthermore, in [3], it is noted that it has been empirically demonstrated that the topology of the network and the position that each bank occupies have an effect on the mechanisms of stress propagation in a crisis. As promising as the subject is, the main obstacle to the development of it, however, besides a certain skepticism, is the availability of data on interconnection between banks. Often, in fact, this information is hidden, for reasons of confidentiality, even if there are methods to avoid revealing to whom a certain piece of data refers.

To strengthen their argument regarding the use of complexity theory techniques in the financial field, the authors indicate an example that investigated how these innovative tools can be used to discover new aspects related to the 2008 crisis and especially the possibility of identifying early warnings (17). The objective of this article was to study the changes in the world trade, in particular the effects of the subprime crisis. To do this, the tool used was a bipartite graph representing the World Trade WEB (WTW)[1]. In this case, the WTW was represented as a network undirected and binary, where the nodes of a layer were countries, the nodes of the other were products. Therefore, a node representing a country was connected to a node that represented a product if the country exported an amount of it greater than a certain threshold. The time window of this analysis was 1995-2010 and allowed to study what happened before the crisis and after it. Bicliques were used to conduct this analysis, i.e. complete subsets of bipartite graphs, where every node of the first (sub-)layer is connected to every node of the second (sub-)layer. Therefore, bicycles of different topologies were used and their evolution over time was studied. What one noticed when considering the abundance of the different bicliques over time was the presence of four trends in the series and that in 2007 what could be described as a structural break occurred. This proved that 2007 was the critical year of the onset of the effects of the crisis, which therefore also immediately affected world trade.

[1] A graph is bipartite if it consists of two layers of nodes and there are no links between nodes on the same layer.

Although interesting, this assertion could only be made in retrospect. What the authors tried to understand was whether it was possible to detect early warnings before 2007, that is whether it was possible to notice the system's loss of resilience before it went into crisis. On the contrary, on the basis of this preliminary analysis, at that precise year the trend was upwards, so it would even have been reasonable to assume that the system was in healthy conditions. What was then done was to assess whether these four trends were statistically significant. In order to perform this analysis, the authors resorted to building a null model and comparing the real network under investigation with it, a classical method to study the specific properties of a complex system in network science.
In practice, what was done was to measure the statistical significance of the discrepancy between the observed values in the real WTW network and the expected values derived from the null model and interpret this measurement as possible changes in the structure of global trade. This more precise assessment showed that global trade in the years considered followed two distinct trends, the turning point of which was 2003. Until the latter year, in fact, the level of statistical significance was more or less constant, after which there was a structural change in global trade that was made evident by the rate of increase in randomness that started four years before the burst of the crisis.
During the second period, 2003-2010, the WTW became increasingly similar to the null model, showing a loss of internal structure, and, after 2007, stabilized. This analysis suggests that the system's loss of resilience occurred during the period 2003-2007 and that this was the crucial period for the developing of the 2008 crisis. On the other hand, 2007 was the year of the unforeseeable event that triggered it and therefore only the final year of an underlying change in the structure of the world trade that started in advance. If the structure of world trade at that time had been looked more closely at, one might perhaps have noticed some early warnings useful to better prepare for the crisis or even to try to contain it. From an economic perspective, the authors interpreted this empirical evidence as the gradual alignment of exports of emerging economies with those of advanced ones. This would have made the system more fragile and thus facilitated the occurrence of a systemic event. The investigation continued by considering several nodes subsets. First, subsets of products were selected, the so-called "macrosectors", and then some groups of countries (e.g. BRICS, G7,..). Again, it appears that there were early warnings that could have been taken into account. In particular, some specific subsets emerge to be more sensitive to changes, e.g. the emerging economies with respect to the advanced ones, and therefore to be useful indicators of a loss of resilience of the system.
This paper proves that tools of economic complexity, in particular networks, can be successfully deployed in the study of economic-financial matters. This observation led to this study of the systemic risk of the European banking sector during the two-year period 2021-2023. In this case, the analysis was carried out with Bayesian Networks. In terms of the results obtained, it was first tried to estimate a single Bayesian Network over the entire two-year period, also making an attempt to consider that financial stress contagion among banks may occur with a few days delay. Once it was found that with only one Network too much information was lost on how much perceived risk changed over the period and that the methodology for accounting for the time lag between spread

growth simply had reduced volatility, two Networks were estimated on the two different years. Despite an improvement in the interpretation of what happened, again no satisfactory conclusion could be drawn. At this point, as a last attempt, the rolling window method was used to estimate 19 half-yearly Bayesian Networks. The resulting estimates show an increase and a concentration of systemic risk in the two years since the beginning of the war in Ukraine. This perception persisted for about four to five months and then declined steadily. Among the banks in the sample, those most responsible for stress amplification were found to be *BNP Paribas* and *Danske Bk A S*. Beyond the results obtained, for which the simplifications made must be taken into account, and the specific tool implemented, which, however, already has its basis existing literature (18), the goal of this section was to demonstrate that there are opportunities in using non-traditional techniques in the economic-financial fields.

Chapter 2

Data Analysis

2.1 The structure of raw data and data preprocessing

The data involved in the analysis were collected from the Markit database, a frequently quoted source when working with CDS spreads (16). Markit is a company that provides daily data in a *.csv* format, so in a tabular form. The data provided are very diverse and include different sectors and countries. It is therefore possible to make different choices and indeed it is necessary, based on one's own technological limitations, to make assessments on which information to base the analysis. It is essential to rely on the documentation provided by the company to map the choices made onto the data, as it is not always straightforward to understand what certain field values refer to. In the remainder of this section it is presented how the raw data were structured and how they were preprocessed accordingly.
Thanks to the documentation provided (12; 14), for each year in the sample (2001-2023) it was possible to clearly understand the information provided and to work only with relevant one. Here it follows a brief description of only the fields and the corresponding values taken into account to carry out the analysis:

1. *Date*: the reference date for the data collection. Since the objective of this analysis was to study the evolution of the systemic risk over a certain period, the time dimension was fundamental. It is important to underline that in the data the dates were provided in the American format, so first the year, then the month and finally the day (YYYY-MM-dd)[1];

2. *ShortName*: the name of the reference entity (chapter 1.2);

3. *Sector*: the Markit Industry Sector of the organization. Regardless of how the different sectors were classified, since the work was based on banks, the analysis was carried out considering only the Sector called *Financials*;

[1]This format convention will remain in use for the remainder of the paper whenever a date is referenced.

4. *Region*: the Region of the organization. The only value considered for this field was *Europe*, as the idea was to work only with European banks. It is important to state that another possible value, *E.Eur*, was excluded because there were no banks taken into account for the analysis. In fact, in the end, a bank was considered European if it belonged to the European Union, Switzerland or United Kingdom. Russia was excluded because it was considered less interconnected with the rest of Europe and would probably have resulted in a big outlier after the invasion of Ukraine;

5. *Currency*: the currency used while issuing the instrument. It is important to take it into account when data are aggregated because currencies have different values and perceptions of risk. For this reason, this field was included in the initial dataset because one cannot mix data referring to different currencies without appropriate adjustments;

6. *RunningCoupon*: the fixed coupon level of the instrument (chapter 1.2);

7. *PrimaryCoupon*: a boolean value that indicates if the given coupon corresponds to the primary curve, i.e. if that fixed coupon, for example 100 bps, is the usual one used in the curve for a CDS of that reference entity. This field is important because CDS with a value of 'No' in this column need further adjustment;

8. *Tenor*: the duration of a CDS contract (chapter 1.2). The analysis was carried out only with CDS with a maturity of 5 years because they are the most liquid and frequently quoted part of the credit curve (5);

9. *ConvSpread*: the conventional spread associated to the contributed CDS curve (chapter 1.2).

Any other field was considered superfluous.

As stated above, the data were collected from 2001 to 2023, but for 2023 the data stopped at 28 February. The process of retrieving the data is not described in detail, as it is considered out of the scope of this analysis, anyway an attempt was made to try to reduce unnecessary burdens with regard to subsequent data management. However, it is important to emphasize that what has been performed is not necessarily (and probably) the most efficient way of doing this. In general, the aim of this analysis was not to apply the best computational techniques for handling information, but completely inefficient operations were avoided. The end result was a dataset with 1260174 rows and 6 columns, namely the dates, the featured financial institutions as reference entities, the currency, the column *RunningCoupon*, the column *PrimaryCoupon* and lastly the spreads. Every tuple therefore referred to a single CDS issued for mitigating the risk of default of these institutions.

At this point, it was necessary to filter the data by selecting only those banks for which systemic risk was to be studied. The European financial institutions with at least a referred CDS with a 5 year Tenor in the data were 159. This number was unfeasible to work with and also was composed of financial institutions that were not banks. The problem therefore arose of identifying a criterion on the basis of which to select banks that could be representative of the entire European banking

system. In order to solve it, the S&P Global ranking of *Europe's 50 largest banks by assets* of the year 2022, thanks to which it was also possible to date back to 2021, was considered (19). Since the data available for 2023 were partial, it was deemed useful to rely only on the previous year's ranking. Therefore the 30 largest European banks were selected, a number that was neither too small to obtain a robust analysis nor too problematic to pose computational problems[2].

It was decided to adopt the criterion *by assets* to rank the banks by size and to select the largest banks as done in [15]. Another possible criterion to rank banks is that of by market capitalization (7), whereas the reason why the largest banks were chosen is that they are expected to be the most decisive in assessing the severity of systemic risk in the financial system. The resulting complete list of featured banks is reported in the Appendix. It is important to emphasize that *La Banque Postal SA*, the 16th largest European bank in 2022, was not present in the data and that *Sberbank of Russia*, which is included in the ranking, is Russian and therefore excluded due to upstream decisions. The number of variables was therefore immediately reduced to 28, a value that was nevertheless considered satisfactory[3].

At this point, once it was ascertained that there were no missing data, the column *Date* was converted to a date format, whereas all the others had the expected object type. This made it possible to find out whether there were CDS referred to all banks for each day of the reporting period. Three institutions did not fulfill this requirement: *Groupe BPCE* and *KBC Group NV* had too few days with associated data, 53 and 75 respectively with respect to 558 of the other banks, to be able to think of any solution and were therefore discarded, whereas *Nationwide Building Society* was held as it had only one day less than the others without CDS. It was therefore possible to include the missing data with a few adjustments without affecting the representativeness of the data, but at this point further analyses were considered needed before deciding which technique to use.

Other choices were also made regarding data preprocessing. First of all, the column *PrimaryCoupon* was filtered only by the value *Yes*, because otherwise further considerations would have to be made and this complication was deemed to be outside the scope of the analysis. This decision also made the column *RunningCoupon* non-informative as all corresponding values were 0.01 and so superfluous. Finally, it was decided to consider only CDS issued in *Euro*, the most prevalent currency, since they were evaluated sufficiently representative and also the region under observation partly overlapped with the Eurozone. All these simplifications obviously had an impact on the robustness of the results obtained, however, given also the limited means to conduct the analysis, particularly of computational performance, and the research question, what was done was considered a good compromise. As a final preprocessing step, the data were aggregated to obtain a single spread value for each bank for each day. The average spread was then multiplied by 10000 in order to work with basis points.

[2]The constructed network in [18], composed of about 20 nodes, was considered as a benchmark.

[3]In general, even during the rest of the analysis, all exclusions deemed necessary to avoid estimation biases and to stick to the research question were made. Once the final dataset was obtained, the number of banks selected was considered sufficient to obtain a robust analysis, even taking into account the expected waiting time for calculations with the available means.

2.2 Exploratory Data Analysis and Time Series Analysis

The Exploratory Data Analysis, shortly EDA, is the usual first step in Data Analysis (6). The aim is to get an initial idea of the available data, their general properties and to detect any special cases. It is also useful for evaluating possible models to be implemented and for creating expectations or hypotheses about future results. On the other hand, Time Series Analysis, shortly TSA, is the discipline that deals with the study of time series (8). A time series is a process that is sequentially ordered over time. The study of this field has two main objectives, at least in the field of finance and economics: the modeling of phenomena and thus the analysis of causal relationships, as in the case of this work, and forecasting. The time series of Credit Default Swap spreads referring to selected banks are defined as univariate, i.e. the expression of a single variable and its study through the realization of its past values. Figure 1 visualizes all the time series together.

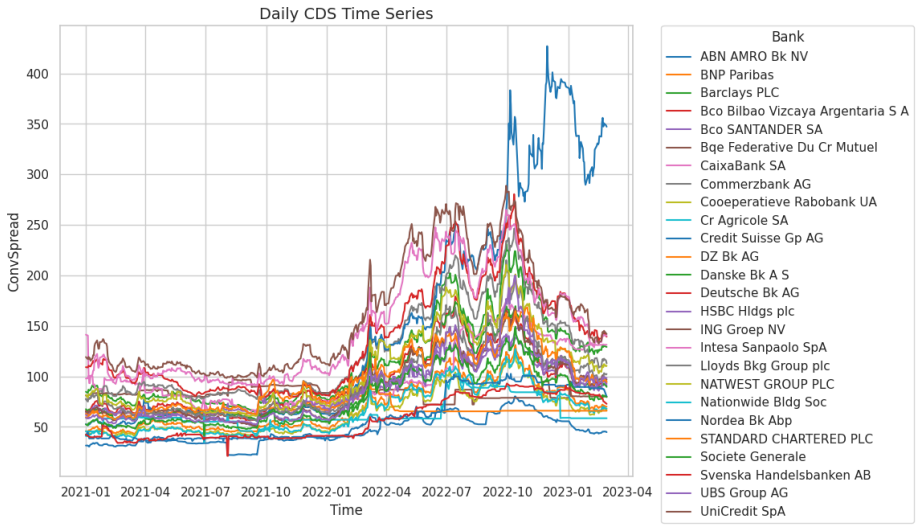
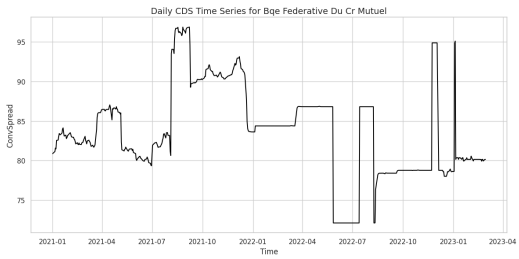


Figure 1: The plot of all the times series of the CDS referred to the banks in the reported period.

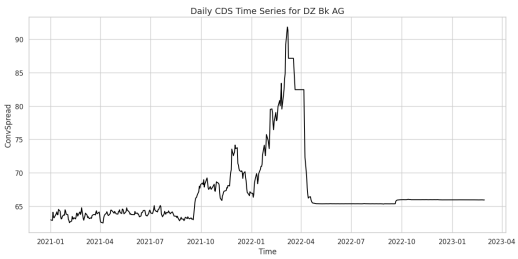
First of all, the series seem to follow a similar behavior, with some exceptions. The spread values were more or less constant from the beginning of the period until January 2022, after which there was an ascent until about August with some peaks, one when the invasion of Ukraine by Russia took place, and some descents, then the spread values fell and rose again until the October spike, after which, while for the other banks the series returned to about the values of January 2021, the spread of *Credit Suisse Group AG* went up again and remained stably higher. Certainly, therefore, the latter could be considered an outlier, at least from October 2022 until the end of the period. It can be assumed that this behavior is due to the first signs of the crisis that hit the bank and led to its takeover by *UBS Group AG* in March 2023 (21). This event anyway did not occur during the period under review and therefore it would have been wrong to take it into account. In general, it is normal for a bank to have a negative period and rather the crisis experienced by *Credit Suisse Group*

AG was an important event to study the systemic risk of the European banking sector between 2021 and early 2023.

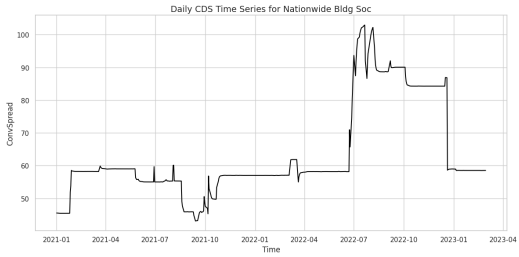
However, other data worthy of attention emerge from the plot, particularly after July 2022. For this reason, since the number of banks considered was not prohibitive, it was decided to visualize each individual series in order to identify any peculiar behavior. There were three banks that for the entire period or for a limited but relevant window behaved very differently from the others: *Crédit Mutuel Group*, *DZ Bank AG* and *Nationwide Building Society*, as it is possible to note in Figure 2, where *HSBC Holdings PLC* is added as a representative of all the other banks. In the three series the two main problems are that there are periods when the spread value was basically linear and abrupt jumps in a very short time, two features absent in all other series. Given the nature of these phenomena, it is conceivable that there could have been some problems with data collection. This could have invalidated the representativeness of the data regarding the real perceived risk of the three banks and therefore it was deemed appropriate to discard them, since they could have introduced biases in the estimations[4].



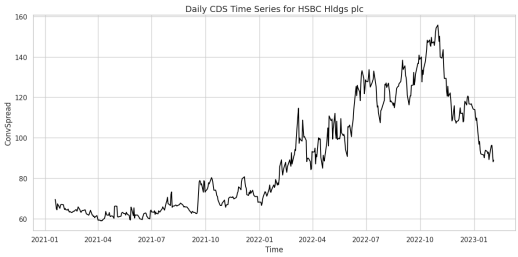
(a) Time series of CDS referred to *Crédit Mutuel Group*.



(b) Time series of CDS referred to *DZ Bank AG*.



(c) Time series of CDS referred to *Nationwide Building Society*.



(d) Time series of CDS referred to *HSBC Holdings PLC*.

Figure 2: The plots of the three discarded banks and *HSBC Holdings PLC* for comparison.

[4]This solved also the issue related to the missing data of *Nationwide Building Society*, specifically which technique was the most appropriate to estimate the value of the spread in order to have the same number of values in the sample for each bank.

At this point, a slightly smaller subset of the period studied so far was selected, i.e. from 2021-02-01 to 2023-01-31, to consider precisely two years. Ten previous and three following days necessary for the algorithms implemented in the following section have been added to the considered biennium. Although there were additional data that were not part of the sample, since they were a very small fraction of the total, it was still possible to obtain and evaluate descriptive data statistics, the second step of the Exploratory Data Analysis. The aim of this part of the EDA was to compare what was being studied with what has emerged in the literature.
First of all, the period under review consisted of 517 days, to which the 13 additional days mentioned above must be added temporarily. The final number of banks considered was 23. The descriptive statistics were obtained by considering together the spreads of all the banks. The average value was around 95.7 bps, with a standard deviation of about 50.58. The minimum value of the spreads was 21.26, whereas the maximum was almost 427 bps. Finally, skewness was about 1.78 and kurtosis about 4.78. The latter two values were in line with what was expected, as time series in the economic-financial field are usually non-Normal (8). The data were in fact leptokurtic, since the kurtosis value was greater than 3, which indicated heavier tails than Normal and therefore a greater chance of extreme events. This value was surely affected by the observations registered in 2022 and confirmed the propensity of banks to go through periods of financial stress, i.e. when spread values are far from the average, jointly. As it is possible to observe in Figure 3 and validated by the value of the skewness, the right tail was longer and heavier. There was therefore a higher probability that spreads would rise rapidly, also because they are bounded below. They in fact cannot be negative and this aspect is not secondary to take into account when analyzing the data. It can also be seen that most values were concentrated to the left of the mean, as described by the skewness.

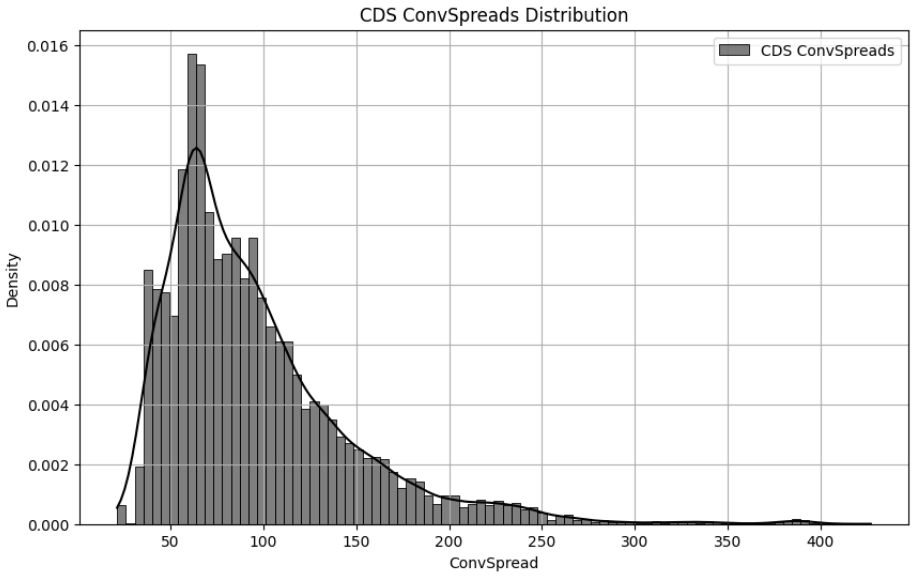


Figure 3: Distribution of the CDS.

After assessing the descriptive statistics of the series, it was useful to study the dynamics of the series as done in [16]. A usually desirable property in a time series is stationarity (8). It requires that the statistical properties of the series do not depend on the period considered and this makes it possible to generalize what was obtained from the sample to the whole series. One way to study stationarity is to consider the first-order autocorrelation, as it is done in [16]. In the 23 series considered, the median and minimum values of the first-order autocorrelation of the spreads were very similar to those found in the reference, since the values recorded were respectively 0.996 and 0.992. These values signaled a very strong persistence, almost at the level of a unit root. It was therefore also useful to apply the augmented Dickey-Fuller test, whose null hypothesis is precisely that there is a unit root and that therefore the series exhibits non-stationary behavior (8). For none of the 23 series this hypothesis was rejected and this implied a random walk dynamic. This result is commonly found also in other financial series, such as interest rates, although in reality, again, it is mainly due to the fact that these series are bounded below, as they do not really behave as random walks (16).

The emergence of unit roots in the series, and thus their non-stationarity, is considered problematic and is usually handled by considering differences between values. In [18], however, this issue is not addressed, to the extent that it is not even checked whether the series meet this assumption, and the model is estimated without the spread values undergoing any transformation. In the present analysis, therefore, the assumption of stationarity is tested and results are compared with those in the literature only to assess how well the data in hand are in line with it, after which the models are estimated without transformations on the data. A more in-depth discussion of this topic is deferred to chapter 5.

Another interesting and important aspect of TSA is how volatility changes over time (8). Usually it is not constant and tends to create cluster of similar values distributed over the whole period. It is therefore useful to visualize in which periods a series display high volatility and in which low, in particular in a financial series. This is also what was detected in this analysis, as it is displayed in Figure 4. Again, all spreads were considered together, so for each day the squared difference between the average daily spread and the global average mentioned above was computed. As expected, the volatility plot resembles the general plot of the data. It in fact increases during periods of instability, when therefore values deviate greatly from the average. This plot shows furthermore how the assumption of stationarity in a time series may prove to be inadequate or otherwise difficult to apply and confirms that the second year is the one in which to expect a higher estimate of systemic risk, given the greater instability.

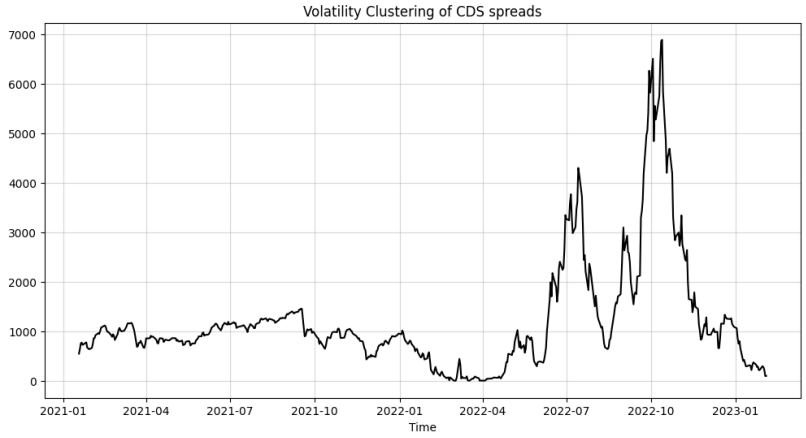


Figure 4: Volatility clustering of the CDS over time.

Another typical aspect of time series, particularly when studying univariate series, is serial correlation, i.e. how variable values depend on their past realizations. It was therefore interesting to deepen this property on the series. To do so, the Ljung-Box test was applied. As done in [16], the log-difference transformation was applied to the series, just for this analysis, and it was tested whether there was autocorrelation up to the tenth lag. In contrast to [16], the incidence of series with significant autocorrelation was 14 out of 23, much less than the benchmark of 98 out of 100, although this was still the majority. Other authors also emphasize the presence of a significant autocorrelated dynamics (5). This is interesting because it shows that the information in the credit market, analyzed from this point of view, is strongly persistent.
The Exploratory Data Analysis and the Time Series Analysis are important steps in the study of historical data. The first allows to visualize the data and thus create expectations about the results to be obtained from the estimation of the models. Systemic risk can be expected to be greatly higher in the period between January 2022 and January 2023 with respect to the rest of the series. It was also useful to see that no major differences emerged in the general behavior of the banks, apart from *Credit Suisse Group AG* after October 2022. Finally, thanks to the Exploratory Data Analysis, three banks whose series showed peculiar features that could have affected the correct estimation of the models were discarded. On the other hand, the Time Series Analysis was useful for assessing the consistency of the data under scrutiny with what has emerged from the literature so far. In general, the series under review were in line with expectations. The next and final section describes how the data was processed to be the input of a Bayesian Network.

2.3 Data transformations for estimating a Bayesian Network

The algorithm implemented to construct Bayesian Networks was based on the concept of *modified ϵ-drawup* (18). This notion allows for the detection of significant jumps within a time series and is therefore used to identify moments of strong market-perceived stress for the reference entity, as the CDS spread has risen rapidly and by a large amount. Once these moments have been identified for each bank, then, it is possible to model the dependency between them based on how often they share a period of rising spreads. It is therefore assumed that if the ascent occurs jointly, then there is a dependency between the entities.

Compared to other methods, it has certain advantages. First of all, there is no need to make any assumption about the distribution of the data on which it is applied. Often some methods require precise distributions, so approximations are sometimes used to apply them. For example, the 'classical' assumption is that the data follow a normal distribution, which cannot be applied to CDS spreads. The modified ϵ-drawup method, on the other hand, is applied directly to the series, whatever it is and whatever properties it has, even if it is very skewed, for example. Secondly, it is scale-free. This means that it is not affected by the unit of measurement and generally by the scale of the data. The output in fact is the number of drawups detected and thus a dimensionless quantity. This makes it possible to compare even very different series. Finally, the identification of a significant jump is done by considering the volatility of a narrow period just prior to the change under investigation. In fact, standard deviation is used as a measure of local variability to set the threshold on the basis of which it is decided whether or not a certain jump is a modified ϵ-drawup. This allows the drawup to be contextualized within the series and the specific period in which it occurs, since, as noted in the Time Series Analysis part, series volatility tends to change even by a great deal over time.

To better understand how this method works, it is useful to trace its origin and the changes it has undergone over time based on the goals that the authors who used it set in the different papers. The modified ϵ-drawup method originated in fact from the symmetrical concept of the *ϵ-drawdowns* (9), since the first application context was the stock market, where the risk is related to a sudden drop in price. Initially the idea was therefore to study drawdowns, i.e. loss from the last local maximum to the next local minimum, in risk assessment of investment strategies, since a more accurate estimate of risk could be derived from the study of this concept. In fact, according to [9], traditional methods, such as variance, do not allow for the correct assessment of risk, especially of a large collapse. In the article the author pointed out that it was necessary to define this notion precisely, as a minimum and a maximum could be detected differently according to different needs. It was then decided to establish a threshold ϵ on the basis of which was determined whether or not a descent was a drawdown, whose definition thus became a descent from a local maximum to a local minimum within which there was no rise above ϵ. Clearly the main disadvantage of this definition is that it is not obvious to establish what the ideal threshold value should be. The development of a sophisticated method for this objective however was deferred to future work, although it was already

assumed that it should have been associated with the volatility of the series under consideration. Once this concept was applied to CDS (10), it was obviously reversed, since the risk in this case is related to spread growth. Thus, the method of ϵ-*drawup* was used to estimate the probability of co-movement among the different institutions considered. In fact, the authors noted that it is difficult to use a network to assess the systemic risk in a financial environment, since the topology is crucial and to estimate it is necessary to model the dependencies among the nodes. Correlation analysis, the technique used for this purpose, has the limitation that, if the correlation between two entities is zero, this is not a sufficient condition to declare that they are independent. This is why it is useful to resort to the ϵ-drawup and generally so whenever it is not possible to observe the links in a network but the dynamics of the nodes reflect the dependency structure. The definition used by the authors was that of a persistent upward movement in a time series until a peak has been reached, after which the time series declines by more than an amplitude ϵ.

The algorithm used to identify these movements consists of several steps. First, all local extremes are detected in the series. Therefore, the first local minimum is considered and it is called candidate. The subsequent pair of maximum and minimum is considered and the difference, called in [10] correction amplitude, is calculated. If it is greater than the local variation recorded between the maximum and ten days earlier, then the movement between the candidate and the maximum is a ϵ-drawup. Precisely, the day of the local maximum is recorded. Otherwise, the algorithm goes on by considering the next pair of local extrema and the same candidate. In this way, the authors proposed a way to set the threshold empirically and related to the volatility of the series. The idea is that if the descent between maximum and minimum is greater than ϵ it means that the previous jump is locally relevant and is therefore a signal of entity stress at that time. Clearly, however, there remains the problem of how to identify how many days it is appropriate to consider when calculating local variation. The authors considered a range of values from 1 to 100 and tried to calculate drawups in 50 series with each value. Given that at the two extremes either all maxima are validated as drawups or too many are ignored, according to the authors it is possible to determine which value is the most appropriate just by visualizing the result. In their case ten days was the best value. This choice was then accepted also by subsequent papers, partly because it was evaluated precisely on the spreads of some CDS.

Finally, a further and final improvement has been made to this technique (1). In this case, the scope was the Portfolio Credit Risk Models and the goal was to enrich a risk estimation model. The authors in fact, in addition to systematic risk, the one due to some common underlying factors, wanted to consider contagion risk, the one whereby the stress of one entity affects another. The main shortcoming of the previous model was the underestimation of real risk and to estimate the contagion risk the authors constructed a network based on CDS spreads. They resorted to the methodology of the drawups but this time the definition of it was an upward movement between a local minimum and a local maximum greater than the standard deviation recorded between the day of the local minima and ten days before. This time each pair of extremes is considered individually and in case it is a relevant drawup it is the day of the minimum that is recorded. The final version

is called modified ϵ-drawup. An example is reported Figure 5, where the first minimum is registered as such, the second is not.

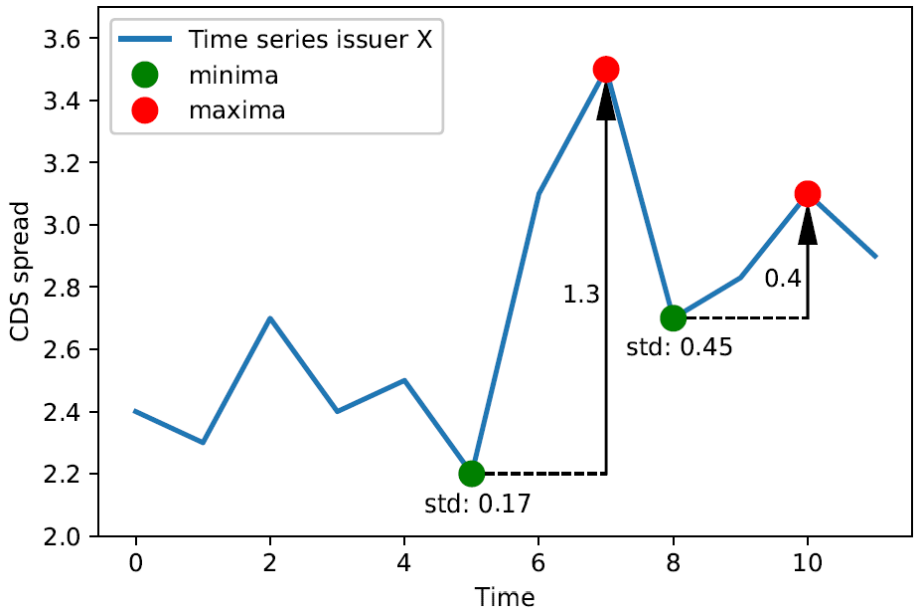


Figure 5: A stylized example of a modified ϵ-drawup from [18].

This brief overview clarifies the reasons behind the choice to use the technique and also illustrates why the 10 days preceding the period under review were included; otherwise the presence of a drawup at the beginning of the series could not have been verified[5]. The algorithm was applied to each series starting from 2021-02-01. The structure of the procedure as applied to the data consists of several steps. The series of the first bank is considered. Starting with the first day of the sample, the algorithm looks for the first local minimum, defined as a day for which both the previous and next corresponding values of the spreads are higher. Once found, it becomes a candidate. At this point the search for the maximum starts from the following day of the minimum. The former has surely a corresponding value of the spread higher than the previous, since the previous is the day of the minimum. Thus, it is sufficient to find the first descent between two days following the minimum to identify a local maximum. If it happens, the two local extremes are registered as a candidate pair. At this point, from the day following the maximum, more minima are sought until the series ends. If otherwise no maximum is found, no pair is recorded and the algorithm proceeds to the second part with the list of the candidate pairs.
This code is designed to meet implicitly certain constraints needed to correctly identify the drawups. In each pair, the maximum value must temporally be subsequent to the minimum. There can be no temporal overlap between pairs and a candidate minimum is discarded if is not followed by a maximum because the series is over. These rules can be broken if a series, even for a short period,

[5]In the following, it will be clarified why it was necessary to keep the 3 days after the end.

has a linear trend interspersed with a few upward (or downward) spikes. This possibility emphasizes the importance of visualizing the data as the first step in the Exploratory Data Analysis, since it allows to hypothesize any aspects to be taken into consideration and evaluate the extent of it. The rules can be broken also if the first extreme of the series is a maximum or the last is a minimum. Given these case histories, it is not enough to simply look for minima and maxima and then pair them.

At this point, clarified the first part of the algorithm, for each pair the difference in spread values between the maximum and minimum is calculated and compared with the standard deviation recorded between the day of the minimum and the previous ten days. If the difference is greater the movement of the series is relevant, as it deviates from the average change recorded over the period, and the day of the minimum is recorded as modified ϵ-drawup, otherwise the following pair is evaluated. Once all the candidates are assessed, the algorithm goes on with another bank until completion. The final result was that for each day and bank in the corresponding column the value *1* was entered if there was a modified ϵ-drawup in the series on that day, otherwise a already prepared NaN remained. Thus, the 10 days prior to the sample were finally discarded as having become superfluous. As the final step, the cells were completed inserting the values *0* or *0.5*, following what was done in [18]: for each bank 0.5 was entered in the cells if at least one of the next 3 days had a modified ϵ-drawup, otherwise 0. The objective of this operation is to take into account the delay in stress propagation between banks. In fact, with this transformation it is possible to consider both cases in which two banks experience a drawup at the same time, so on a certain date the value of the corresponding cell is 1 for both, and the case in which the drawup of one bank temporally follows the drawup of another in a limited window of time.

Once the additional 3 days at the end of the sample were exploited for this last operation, they were dropped. Once the operations on the dataset were completed, it was possible to assess the relative frequency of a modified ϵ-drawup for the various banks. On average, on just over 63% of the days no drawup was recorded. On the other hand in about 9% of the days it occurred and finally in almost 27% the value of 0.5 was recorded. The banks least aligned with these values were found to be *ABN AMRO Bank NV*, with almost 78% of days without a modified ϵ-drawup, the highest value recorded, whereas on the other hand *CaixaBank SA*, *HSBC Holdings PLC* and *Nordea Bank Abp* had all values lower than 60%. All the others had values between 60 and 70%.

The series whose spreads were used to estimate the models is shown in the Figure 6.

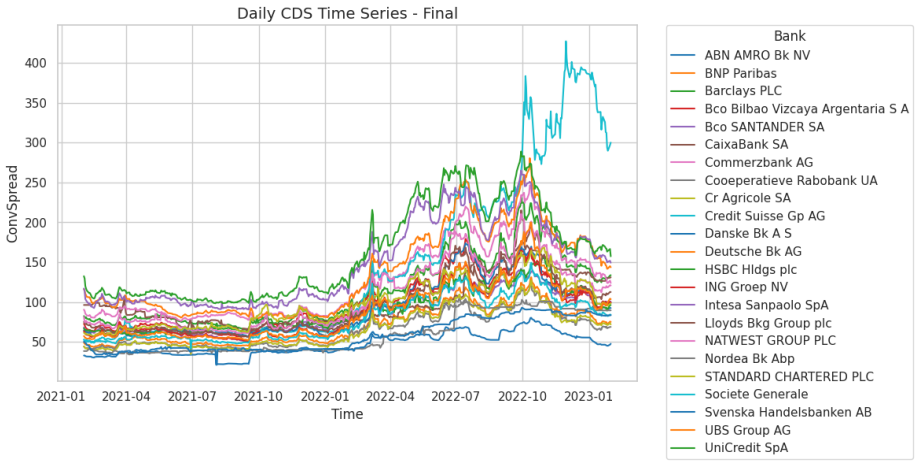


Figure 6: The plot of the times series of the CDS referred to the banks in the sample period.

In the next chapter the model implemented to estimate the systemic risk of the banks, Bayesian Network, is presented.

Chapter 3

Statistical Modeling

3.1 Probabilistic Graphic Models

Studying the systemic risk of the banking sector means considering an event during which a loss occurs in a large proportion of the banks in the system. The goal of this analysis is to assess how the likelihood of such an event has changed over the considered time window. In order to obtain a plausible result, it is essential to properly model the interactions of dependence among banks in the system. In other words, one must derive from the data how likely is that a bank struggles to repay its debts given that another bank is in that condition. This concept is important for the study of systemic risk because if banks are highly interdependent on each other, it is more likely that the financial stress of one will be followed by that of others. In addition, it is important to understand which banks are more likely to trigger the cascade effect. Once the dependencies among the entities in the system have been characterized, the joint probability distribution can be defined, i.e. the probability distribution of all combinations of the possible values for each bank. When the joint probability has been constructed, it is possible to derive the probability that one or more banks are under stress given that others are or are not, and therefore assess the systemic risk. It is easy to realize that, even if for each entity the set of possible assignments is 2, the total possible values of the joint distribution is equal to 2^n, where n is the number of banks in the system. Specifying a distribution whose number of values follows an exponential function requires appropriate tools, as it already becomes intractable for a not particularly large number of banks, such as in the case of this analysis[1].

Some tools to handle this complex distributions are part of the family of *Probabilistic Graphic Models* (11). The name already describes the two main features of these models. First, as already illustrated, they are useful for building and even representing probability distributions of complex systems, where therefore the number of variables considered is high and there are dependencies among them. The notion of dependence is central to this field and requires the definition of certain

[1] Precisely, $2^{23} = 8\,388\,608$ possible values.

concepts that belong to probability theory. The probability of whether or not a certain bank is under stress given the observation of another bank's status is called conditional probability, and is defined as, given two events α and β:

$$P(\beta \mid \alpha) = \frac{P(\alpha \cap \beta)}{P(\alpha)}. \tag{4}$$

From Equation (4) it follows the *chain rule* of conditional probabilities. In fact, if generalized to more than two events, it is possible to state that, given k events from α_1 to α_k:

$$P(\alpha_1 \cap \alpha_2 \cap \cdots \cap \alpha_k) = P(\alpha_1)P(\alpha_2 \mid \alpha_1)P(\alpha_3 \mid \alpha_1 \cap \alpha_2) \cdots P(\alpha_k \mid \alpha_1 \cap \cdots \cap \alpha_{k-1}). \tag{5}$$

Based on Equation (4), it follows the *Bayes' rule*:

$$P(\alpha \mid \beta) = \frac{P(\beta \mid \alpha)P(\alpha)}{P(\beta)}. \tag{6}$$

These definitions hold even if, instead of considering single events, we consider random variables, that is, functions that associate a variable with a set of events through probabilities. The probability distribution of the variable with respect to the possible events is called the marginal. It is also possible to characterize the distribution of probabilities of a variable according to the events of another variable, namely the conditional distribution. Once the former is derived, the subsequent rules can be stated as for the single events. In this analysis, each bank corresponds to a random variable that can take on certain values indicating whether a modified ϵ-drawup occurred, which signals a relevant growth of spreads and so a situation of financial distress.
Once these notions are defined, it is possible to introduce the concept of independence between two banks, for which one variable X is independent of another variable Y if the conditional distribution of X given Y is equal to the marginal of X. In other words, observing Y provides no information about X (and vice versa). However, it is difficult for two variables to be completely independent, while instead a less restrictive property, and thus more realistic, is *conditional independence*. The idea behind this property is that X is independent of Y once conditioned to a third variable Z. This means that Z contains all the information needed to describe the distribution of X and therefore observing the realization of Y adds no information. As much as X and Y are not independent, both become so once conditioned to Z, since conditional independence is a symmetric property. Intuitively, this property is desirable because it allows to reduce the possible combinations to consider for specifying the joint probability values.
The second feature of the Probabilistic Graphic Models is the use of graphs. Each of these mathematical tools consists of nodes, which represent the entities in the system, and edges, which represent the interactions between the nodes. Edges can be directed, i.e. that goes from a node to another and indicates the direction of the interaction between them, and in this case they are usually depicted as arrows, or undirected, where there is no direction in the connection. Hence, directed and undirected

graphs are differentiated, as it is usual to consider the two possibilities mutually exclusive. In the first case, the node from which the arrow starts is called the parent and the node to which it reaches the child. Starting from a node and following the edges according to their directions results in a directed path. A cycle is a directed path for which the finishing node is the same as the one from which it starts. These concepts are important in defining a particular type of model: a directed graph for which no cycles exist, namely a *directed acyclic graph* (shortly DAG). The Probabilistic Graphic Models that are implemented in this analysis use a DAG to encode a complex probability distribution. The nodes represent the variables and the edges the probabilistic interaction between them. A simple example is reported in Figure 7.

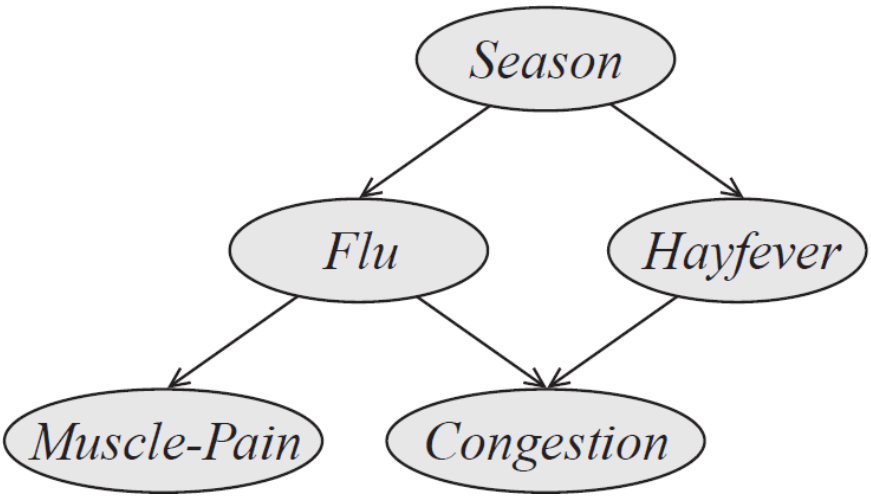


Figure 7: A simple example of a Probabilistic Graphic Model from [11].

First, one is able of understanding which variables are conditionally independent. In fact, the variables *Congestion* and *Muscle-Pain* are conditionally independent given *Flu*, because there is no arrow between them and they have a common parent. As part of this analysis, this model allows to represent the graph of the 23 banks and draw conclusions regarding the dependencies among the various entities. Second, the joint probability is represented in a compact way, as it is "broken" in factors. In fact it is possible to write that[2]:

$$P(S \cap F \cap H \cap C \cap M) = P(S)P(F \mid S)P(H \mid S)P(C \mid F \cap H)P(M \mid F), \tag{7}$$

which is called factorization and it requires fewer parameters than the one related to the exponential function.
It is possible to state more precisely the advantages of this family of models so that it is easier to compare them with other solutions. From the point of view of representation, as has already been shown, it makes it possible to derive information immediately about the relationships between the

[2]For simplicity, variables are indicated by the first letter of their names.

variables and allows the distribution to be written in a tractable way even when n is large. It is also intuitive and easy to understand and evaluate for those who are not insiders. Secondly, a model can be built using a purely data-driven approach, with which, however, the human experience can be associated, for example in constraining the model to include certain dependencies. Finally, it also makes it possible to perform inference and thus derive the probability of a certain event given observations on the other variables. These advantages related to these three basic components (representation, learning and inference) make this family of models an excellent tool for solving a wide range of problems. An in-depth section is devoted to each of these to understand how from the data it is possible to represent the model and draw initial considerations, learn the most suitable topology and parameters, and finally make inference.
Lastly, a *Bayesian Network* is a probabilistic graphic model that makes use of a directed graph (as in Figure 7), while Markov Network makes use of an undirected one. What has been generally stated about Probabilistic Graphic Models so far applies to both, but from now on the rest of the chapter is devoted to the first model, as it was the one actually used in the analysis.

3.2 Representing a Bayesian Network

Consider to represent the joint probability distribution of n variables. Even given the simplest case, whereby each variable can take only two assignments, the total number of possible values is 2^n, and $2^n - 1$ parameters are needed for the complete specification of the joint distribution. There are several reasons why, even for a small value of n, the representation of such a joint probability is challenging. First, computationally it is difficult to manipulate and expensive to store in memory, so management is often time-consuming. Second, a human can hardly handle so many numbers and derive information from them. Often, moreover, some of the resulting probabilities are so small as to be meaningless. This fact can be tied to a combination of events that are difficult to achieve, for which therefore there is also little interest, but which nevertheless must be taken into account for specifying the joint. Finally, robust estimation of such a large number of parameters requires an immense amount of data.
In order to represent a joint probability distribution, the concept of independence and its properties can be exploited. For simplicity, consider the case of n binary variables. The simplest case is one for which the variables are marginally independent of each other (the usual example is tossing coins). In fact thanks to the chain rule, whose formula is stated in Equation (5), it is possible to define the joint probability as the product of the individual marginals and, consequently, the total number of parameters to be estimated is n. However, this alternative parametrization, which is much more manageable, often proves inapplicable because in real cases it is hardly possible to assume that the variables are marginally independent.
Once the possibility of assuming marginal independence between the variables is discarded, a more compact representation can still be derived. Consider three variables: I, S, and G. If, for example, the first two are binary and the third has three possible assignments, the joint probability has 11

parameters and 12 possible values. At this point an alternative representation of the joint can be derived if a conditional independence property is met, that is, if, for example, G and S are conditionally independent given I. This hypothesis implies that I is the only reason why G and S are correlated and therefore there are no other factors involved. It is clear that the plausibility of this assumption depends on the context, however evidently in absolute terms this statement often turns out to be "false" and only an approximation. However, if this approximation is considered acceptable, then it is possible to derive the joint probability in the following way[3]:

- resorting to the mentioned chain rule:

$$P(I, S, G) = P(I)P(S|I)P(G|I, S), \tag{8}$$

- applying G and S conditional independence given I:

$$P(I, S, G) = P(I)P(S|I)P(G|I). \tag{9}$$

Therefore, based on this assumption, the factorization of the joint probability can be obtained as a product of three *conditional probability distributions* (CPD). This alternative parametrization is more compact than the original one as it requires seven parameters. As the number of variables and thus the complexity of the joint distribution increases, the reduction can be even more pronounced. Another advantage of this parametrization is modularity. In fact, if, for example, one wanted to remove the variable G from the system and consider only I and S, one would simply discard the CPD of G given I and would get the joint of I and S. Vice versa, if one now wanted to add another variable, the process would be mirrored and still simpler, whereas in both cases the original distribution would have to be rewritten all over again.
This simple example demonstrates the importance of the concept of conditional independence. The other fundamental concept for representing a Bayesian Network is the one of DAG. Consider now a more complex scenario, for example the one reported in Figure 7. The nodes of the Bayesian Network represent the random variables and the edges represent the direct influence of one on the other. Each variable is thus a stochastic function of the values that its parents can take, while the edges represent how the system works, i.e., what happens independently of everything else and what needs to be traced back to the assignments of the parents instead. Each variable is therefore associated to a CPD that specifies the values of the variable given all the values of the parents. For those nodes that have no parents, the CPD actually represents the marginal distribution. So the structure of the network, and its consequent dependencies, and the set of local probability models, i.e., CPDs, each of which refers to a single variable, forms a Bayesian Network. The DAG is thus its natural representation. Thanks to this explanation, now it is clearer to comprehend the joint probability distribution of the example, already stated in Equation (7), and especially the hypotheses assumed

[3]For convenience, comma is used to indicate the intersection instead of $\cap$. This convention remains in use from here on.

and their translation into interactions thanks to the graph.
However, a clarification needs to be made at this point. It is incorrect to infer that a variable is conditionally independent of network variables once its parents are given. Indeed, this is true from the perspective of the causal mechanism, whereby the parent's value influences the probability of the child's assignments, and not vice versa. Intuitively, in fact, again considering the Bayesian Network in the Figure 7, if the value of M is observed, the probability of observing a certain value of F changes. It is correct, therefore, to state that once the values of the parents are observed, no information can influence beliefs about the probability of a certain variable, other than from descendants. This reasoning is fundamental in that it allows to state that the graph incorporates a set of local independencies, for which each variable is conditionally independent of the nodes that are not its descendants given its parents.
Thus, on one hand, the structure of a Bayesian Network, a DAG, incorporates a set of local independencies, defined as above. On the other, each node in this graph is associated with a conditional probability distribution, each of which is a factor in the alternative parametrization of the joint probability derivable through the chain rule. The fundamental point is that these definitions are ultimately equivalent: a distribution P satisfies the local independencies associated with a graph G if and only if P is representable as a set of CPDs associated with the graph G.
Define I(P) the set of independence assertions of the type $(X \perp Y \mid Z)$ that hold in P. G is thus an independency map of P (*I-map*), that is, the set of local independencies of G is a subset or is equal to I(P). Formally, defined K any graph object associated with a set of independencies I(K), K is an I-map for a set of independencies I if I(K) $\subseteq I$. It is therefore possible to state that G is an I-map for P if G is an I-map for I(P). This means that all independencies embedded in G must be present in P, while the reciprocal does not apply, so it is possible that there are additional assertions in P. So, since the graph G representing a Bayesian Network incorporates a set of independencies, then they must be satisfied by any distribution for which G is an I-map. Consider any distribution P for which the graph in Figure 7 is an I-map. According to the chain rule:

$$P(S, F, H, M, C) = P(S)P(F|S)P(H|S, F)P(M|S, F, H)P(C|S, F, H, M), \tag{10}$$

which does not require any assumption, making it possible to apply the assertions of conditional independence. If the corresponding graph is an I-map for P, then, for example, $(F \perp H \mid S) \in \mathcal{I}(P)$, and in general, applying all assertions, the factorization reported in Equation (7) holds. This factorization applies to all the distributions for which G is an I-map. Thus, if G is a Bayesian Network graph for the variables $X_1, \ldots X_n$, the distribution P factorizes according to G if P can be expressed as:

$$P(X_1, \ldots, X_n) = \prod_{i=1}^{n} P(X_i \mid \mathrm{Pa}_{X_i}^G), \tag{11}$$

called *chain rule for Bayesian Networks.* The factors $P(X_i \mid \mathrm{Pa}_{X_i}^G)$ are called the conditional

probability distribution for each variables and correspond to the local independencies. At this point it is also possible to give a proper definition to a Bayesian Network B: a pair (G,P) where P factorizes over G and P is specified over a set of CPDs associated with G's nodes.
Assertions of conditional independencies thus imply factorization. Since the two definitions are equivalent, the reverse is also true: factorization according to G implies the assertions represented. Define G a Bayesian Network structure with nodes $X_1, \dots X_n$ and P the joint distribution of these variables. If P factorizes according to G, then G is an I-map for P. Illustration of the proof of this statement is deemed superfluous and therefore referred back to the source (11), as the focus of this section was to state the equivalence between the two definitions and to introduce to the fundamental concepts. In any case, the reciprocity between the two statements, indicated by the "if and only if," holds.
Finally, the number of parameters to be estimated for a Bayesian Network can be more clearly delineated. The starting point is $2^n - 1$, the number of parameters in the original distribution, which is considered infeasible. Define k the number of parents that at most a node can have. Therefore, the maximum number of parameters to be estimated is less than $n \cdot 2^k$. Even when n is very large, it is possible to assume that k is quite small, as it indicates the maximum number of variables that directly influence a node. Usually one variable is directly affected by a much smaller fraction of variables than the total. It is therefore possible to state that the number of parameters to be estimated for a Bayesian Network is *exponentially smaller* than that of the original joint distribution. This result shows how advantageous it is from this point of view to implement when conditional independence is assumed, which often turns out to be an acceptable approximation of the real world.
Starting from the issue of managing a very complex probability distribution, this section dealt with representation and illustrated the underlying assumptions and the resulting advantages of resorting to a Bayesian Network. The next section illustrates firstly how the parameters of the CPDs can be estimated from the data and how the topology of the network can be learnt.

3.3 Learning a Bayesian Network

In the previous sections the Bayesian Network in Figure 7 was given as an example. The objective of this section is to illustrate how it was estimated. A Bayesian Network can be constructed manually. Once the variables are selected, the network topology, node interactions and the parameters of CPDs (which are not specified in the example but are necessary to define a Network) are established. However some problems arise. First of all, construction requires a great deal of expertise on the subject under consideration, which furthermore can encompass multiple fields of knowledge. In addition, a lot of time is required, and both of these factors may be impossible to satisfy when the number of variables is simply too large. Even if the Network were then actually built, there is often not one that fits every scenario, and this implies additional redesign efforts.
On the other hand, examples of the distribution under consideration can often be obtained. In the case of the example, there may be a dataset that has been compiled and thus can be used to estimate the model. If so, then a model of the distribution of population subjects could be learned. This also does not preclude a subject matter expert from supervising and refining the result based on experience. The task of constructing a model from a set of instances is called *model learning*. Formally, define P the distribution that govern the selected variables and is induced by a directed network model $M^* = (K^*, \theta^*)$. K^* represents the associated graph and θ^* the set of parameters that define the CPDs. The dataset of J examples from P is defined as $D = \{d[1], \ldots, d[J]\}$. The standard assumption is that the instances are sampled independently from P, so they are *iid*. The task is to learn M^*.
Model learning for Bayesian Networks is thus divided into two parts: *parameter learning* (θ^*) and *structure learning* (K^*).

3.3.1 Estimating the parameters of a Bayesian Network

To delve into how to estimate the parameters of a Bayesian Network, it is first necessary to assume the structure of the Network as given and fixed and that the instances of the dataset are observable. This second assumption is less obvious than one might expect, yet it must be assessed before estimating the parameters and, if true, simplifies the process. Define θ the set of parameters to estimate. There are two parameter estimation methods for Bayesian Networks: *Maximum Likelihood* and *Bayesian*.
The criterion on the basis of which Maximum Likelihood Estimation evaluates the goodness of an estimator is how accurately it predicts the available data. Put differently, it is assessed whether the available data are likely given the estimator. The tool for doing this is the likelihood function, which describes the probability of obtaining those instances as a function of the predictor. At this point one looks for the value of the parameters that maximizes likelihood and thus obtains the maximum likelihood estimator (MLE). Usually, log-likelihood is used for convenience, as the logarithmic transformation does not invalidate the search for the maximum.

Consider the simplest possible example for which the two assumptions underlying parameter learning, mentioned above, are met: a network structure consisting of two nodes, X and Y. The two variables are binary and there is a single edge from X to Y. The likelihood function can be thus written as[4]:

$$\mathcal{L}(\theta : D) = \prod_{j=1}^{J} P(x[j], y[j] : \theta), \tag{12}$$

where $x[j]$ and $y[j]$ are the $j-th$ instance of X and Y respectively. Now the interaction between the two variables explained by the network can be exploited. This allows to rewrite the joint probability of the instances as the conditional probability given θ of the instance of X multiplied by the instance of Y conditioned also on X. The final form of the likelihood is:

$$\mathcal{L}(\theta : D) = \left(\prod_{j=1}^{J} P(x[j] : \theta) \right) \left(\prod_{j=1}^{J} P(y[j] \mid x[j] : \theta) \right), \tag{13}$$

which is a product of distinct terms, each for every variable. These terms are called *local likelihood* functions and describe how accurately a variable is predicted by its parents. At this point it is possible to further decompose each local likelihood into a product of other terms, each related to a group of θ parameters. The term for the first variable is more intuitive, whereas, in the case of the second term, it can be further decomposed into two terms, one for each pair of parameters of Y given the value of X. This property is called likelihood function decomposability.
Generalizing, keeping the same reasoning as above, one can start from the general formula of the likelihood function for a Bayesian Network, that is:

$$\mathcal{L}(\theta : D) = \prod_{i=1}^{n} \left[\prod_{j=1}^{J} P(x_i[j] \mid \mathrm{pa}_{X_i}[j] : \theta) \right]. \tag{14}$$

Since local likelihoods are defined as:

$$\mathcal{L}_i(\theta_{X_i|\mathrm{Pa}_{X_i}} : D) = \prod_{j=1}^{J} P(x_i[j] \mid \mathrm{pa}_{X_i}[j] : \theta_{X_i|Pa_{X_i}}), \tag{15}$$

where $\theta_{X_i|\mathrm{Pa}_{X_i}}$ denotes the subset of parameters that determines $P(X_i|Pa_{X_i})$, the likelihood can be rewritten as:

$$\mathcal{L}(\theta : D) = \prod_{i=1}^{n} \mathcal{L}_i\big(\theta_{X_i|\mathrm{Pa}_{X_i}} : D\big). \tag{16}$$

This shows that it is possible to estimate the parameters of a Bayesian Network using MLE by maximizing the likelihood of each CPD independently of the others. This important property is

[4]For convenience, : is also used to indicate |. This convention remains in use from here on.

defined as the *global decomposition of the likelihood function*. This holds only if the parameter sets $\theta_{X_i|\mathrm{Pa}_{X_i}}$ are disjoint, i.e. each variable distribution is specified by a parameter set that does not overlap with the others. This constraint is usually met in the majority of cases.
Assume now that a coin is tossed ten times and that 8 times the outcome is heads and 2 times tails. According to the Maximum Likelihood Estimation method, the parameter describing the probability of getting heads is 0.8. This conclusion would probably be considered hurried, because the *prior knowledge* of the phenomenon, in this case flipping a coin, would suggest that the outcomes are equiprobable and that 10 flips are not enough to conclude that the coin is rigged. On the other hand, however, if flipping the coin 1,000 times resulted in 800 times heads, then perhaps the previous conclusion would seem more well-founded. Prior knowledge therefore is a starting point, but it does not completely invalidate what the data show. To overcome the MLE method, which does not differentiate between the two cases, it is thus necessary to resort to the Bayesian method, which incorporates assumptions prior to observing the data.
In order to incorporate this prior knowledge, the prior distribution of the parameter, that is, the probability distribution assigned to the parameters before observing the data, is defined. Thus, this represents the subjective probability to be assigned to different possible parameter values. At this point, therefore, the posterior distribution can be defined as:

$$P(\theta \mid x[1], \ldots, x[J]) = \frac{P(x[1], \ldots, x[J] \mid \theta) P(\theta)}{P(x[1], \ldots, x[J])}, \tag{17}$$

where the first term in the numerator is the likelihood, the second is the prior over the parameters and the denominator is a normalizing factor. In addition to the prior definition, once the posterior is derived, the Bayesian method further differs from the MLE method since it does not select a single value for the parameters. In the MLE, in fact, the likelihood function is maximized and the resulting parameters are chosen. In the Bayesian method, on the other hand, the entire posterior is used to predict the value of the parameters in the next extraction. It is therefore integrated in the process and thus it is possible to balance the 10 flips in the example and yet conclude that the coin is probably rigged after 1000.
Using the Bayesian method to estimate Network parameters involves deciding which prior best fits the problem and reflects the prior knowledge one has. A first option is the uniform distribution, whereby the possible parameter values are equiprobable. To resort instead to more sophisticated solutions, it is useful first to consider what characteristics a good prior must have. Clearly it is desirable that it can be represented compactly, so for example by an analytical formula. Second, it should allow efficient updating of data. One prior that fits this description is, in the case of this example, the *Beta distribution*. It is characterized by two hyperparameters, α_1 and α_0, real and positive, and a normalizing constant γ that depends on the two hyperparameters and the Gamma distribution.

Formally, the Beta distribution is specified as:

$$P(\theta) = \gamma\theta^{\alpha_1-1}(1-\theta)^{\alpha_0-1} = \frac{\Gamma(\alpha_1+\alpha_0)}{\Gamma(\alpha_1)\Gamma(\alpha_0)}\theta^{\alpha_1-1}(1-\theta)^{\alpha_0-1}, \tag{18}$$

where $\Gamma(x)$ is the Gamma function. Now consider again the example of the coin toss and that θ follows a Beta prior. The goal is to derive the marginal distribution of X, the random variable representing the coin toss, after a single flip. Formally, resorting to standard integration techniques:

$$\begin{aligned} P(X[1] = x^1) &= \int_0^1 P(X[1] = x^1 \mid \theta)P(\theta)\, d\theta \\ &= \int_0^1 \theta P(\theta)\, d\theta \\ &= \frac{\alpha_1}{\alpha_1+\alpha_0}. \end{aligned} \tag{19}$$

This result suggests that the role of the two hyperparameters is to take into account how many times heads or tails were obtained before conducting the experiment. It is like adding an imaginary sample to balance the results obtained from the real sample on which the study is being done. This is precisely why the Beta distribution is a good prior for a binary variable when using the Bayesian method. Indeed, it allows the two hyperparameters α_1 and α_0 to be modeled and thus firstly to establish prior probabilities, and secondly to refine how much prior knowledge should be taken into account and how much data is needed before deviating from it. In fact, $Beta(1,1)$ and $Beta(10,10)$ both predicts the probability of getting heads in the first flip to be 0.5, and thus both consider the two outcomes to be a priori equiprobable, but in the second case more data are required to deviate from 0.5. The Bayesian method therefore, unlike the MLE method, takes into account the two distinct cases from which this reasoning started.

To sum up, MLE involves finding the best parameters, intended as those that maximize likelihood, and thus their point estimates, whereas the Bayesian method requires to keep track of the beliefs about θ. In order to do this, it is necessary to specify the initial uncertainty and then use probabilistic reasoning, i.e. the Bayes' rule defined in Equation (6), to update the distributions. The parameters are therefore for all intents and purposes treated as random variables. First, the joint probability of the data and parameters is defined as $P(D,\theta) = P(D|\theta)P(\theta)$. The first term is the likelihood function and the second is the prior distribution, which captures the initial uncertainty about parameter values and also is modeled based on past experience in the field. At this point the formula of the posterior distribution reported in Equation (17) is derivable by resorting precisely to the Bayes' rule:

$$P(\theta|D) = \frac{P(D|\theta)P(\theta)}{P(D)}. \tag{20}$$

The term in the denominator is called *marginal likelihood* and corresponds in practice to the a priori probability of observing that data given the a priori beliefs.

At this point it is worth investigating when the desirable characteristic for which the posterior is

compactly representable is satisfied. As can be seen from the formula, the form of the posterior depends on the form of the prior. There are priors that are better suited for this purpose, based on the context, such as the Beta distribution for a coin toss. Consider, for example, a problem that includes a multinomial distribution over T possible assignments, for which thus, given T possible outcomes, $\sum_{t=1}^{T} \theta_t = 1$. The formula of the likelihood function in this case is $\mathcal{L}(\theta : D) = \prod_{t=1}^{T} \theta_t^{J[t]}$, where $J[t]$ indicates how many times a certain outcome was observed. Since the posterior is derived by multiplying the prior and likelihood, it seems natural to require the two to have a similar form. In this case, one can resort to the *Dirichlet distribution*. It is a generalization of the Beta distribution described earlier. Formally, the set of parameters $\theta \sim \text{Dirichlet}(\alpha_1, \alpha_2, \ldots, \alpha_T)$ if $P(\theta) \propto \prod_{t=1}^{T} \theta_t^{\alpha_t - 1}$. If the Dirichlet is specified as the prior, then the posterior will also follow that distribution. So in this case not only does the posterior have the advantage of being able to be represented compactly, it even has the same representation as the prior, an even more favorable situation, as computation and representation are further facilitated. In this case a family of priors is said to be *conjugate* to a particular model, as Dirichlet ones are to multinomial and Beta are to Bernoullian ones.
The Bayesian method thus involves a process of updating beliefs about the set of parameters θ. This process begins with the definition of the prior distribution of the parameters and results in the definition of the posterior when the data are observed. Once the posterior is specified, it is possible to study the properties of the model or predict the probability of future experiments. In the case of the Dirichlet distribution, the probability that the next instance is equal to the assignment x^k can be stated as:

$$P(x[J+1] = x^k \mid D) = \frac{J[k] + \alpha_k}{J + \alpha}, \tag{21}$$

where α indicates $\sum_j \alpha_j$. Once again, as in the case of the Beta distribution, the hyperparameters play the role of counterbalancing the observed data in the sample with prior knowledge, as if an imaginary sample were taken into account. For this reason, the Dirichlet hyperparameters are called *pseudo-counts*. The total α reflects the confidence in the prior and is often called *equivalent sample size*.
At this point it is possible to illustrate the estimation of the parameters of any Bayesian Network using the Bayesian method. It was assumed at the beginning of the section that the structure of the Network is given. As explained in the previous section, it incorporates dependencies between the variables. This implies that, in defining the priors of the parameters of individual variables, it is reasonable to assume that they are independent of each other a priori, i.e. observing one parameter does not provide information in the estimation of another, since the interactions between the variables are already captured by the structure. This avoids having to specify complex dependency structures between parameters. This assumption is known as *global parameter independence*. It greatly simplifies the process, but it is necessary to consider whether it is appropriate in the context under investigation.

Define $P(\theta)$ the prior distribution of all possible parametrization of the network, then:

$$P(\theta \mid D) = \frac{P(D \mid \theta)P(\theta)}{P(D)}. \tag{22}$$

The likelihood can be decomposed into local likelihoods:

$$P(D \mid \theta) = \prod_i L_i(\theta_{X_i|Pa_{X_i}}). \tag{23}$$

Moving to the parameters, the global parameter independence is applied:

$$P(\theta) = \prod_i P(\theta_{X_i|Pa_{X_i}}). \tag{24}$$

From Equation (23) and (24), the posterior equation becomes:

$$P(\theta \mid D) = \frac{1}{P(D)} \prod_i \Big[L_i(\theta_{X_i|Pa_{X_i}} : D)\, P(\theta_{X_i|Pa_{X_i}}) \Big], \tag{25}$$

where each subset $\theta_{X_i|Pa_{X_i}}$ of θ appears only once and therefore it is shown that the posterior can be written as a product of local terms. Formally:

$$P(\theta \mid D) = \prod_i P(\theta_{X_i|Pa_{X_i}} \mid D). \tag{26}$$

Decomposing the posterior in this way is useful, for example, to predict the probability of a certain outcome at the next experiment given observations on the posterior, since it is possible to solve this problem separately for each variable. Once solved the problem for each CPD independently of the others, the local solutions can be combined to get the overall one. Additionally, as in the case of the method of the maximum likelihood, it is possible to decompose any local likelihood for each group of parameters of θ. In this case it is stated that a certain prior satisfies the *local parameter independence*. Applying it to the general case, if a prior $P(\theta)$ satisfies global and local parameter independence, then:

$$P(\theta \mid D) = \prod_i \prod_{pa_{X_i}} P(\theta_{X_i|pa_{X_i}} \mid D). \tag{27}$$

If the prior is a Dirichlet, then the posterior is also a Dirichlet and the hyperparameters still plays the role of pseudo-counts when predicting.
In the general case, Dirichlet distributions, characterized by their respective hyperparameters, are associated as priors with each node in the Network. Manually assigning each hyperparameter a certain value is arduous or even impractical. One solution may be to resort to the *K2 prior*, whereby each hyperparameter in the network is assigned the same value. Another idea is to use a prior for which the hyperparameters come from the imaginary sample derived from the prior knowledge. If one had an imaginary dataset D′ of "prior" examples, it would be sufficient to count to derive the

values of the hyperparameters. Clearly, however, this additional dataset should be stored. One could instead store the corresponding probability distribution P′, which could be represented precisely by a Bayesian Network, whose structure can be completely disconnected from the one for which the prior is sought, and infer the probabilities of individual events to derive the value of the hyperparameters of the original Network. The prior to be used to adopt this reasoning is the *BDe prior*.

3.3.2 Learning the structure of a Bayesian Network

In the previous section, to estimate the parameters of the Network, it was assumed that the structure was given. There are cases where indeed the graph is provided, others where it is necessary to build it. This circumstance also involves dealing with parameters. This section will explore how the structure can be estimated from scratch. Again, it is assumed that the data are fully observable.
There are three possible approaches of structure learning. The first is called *constraint-based* structure learning. The idea is to exploit the Bayesian Network definition as a set of dependency assertions. It therefore involves testing to detect conditional dependencies and independencies in the data and constructing the Network that best explains them. The second approach involves generating multiple structures and taking the average of them. These *Bayesian model averaging* methods can be very efficient for certain classes of models. The third approach, the one implemented in this analysis, is called *score-based* structure learning. It involves the specification of a statistical model and the subsequent identification of the best structure. It is therefore a classic model selection problem: define the set of all possible networks that can be a solution (*hypothesis space*) and then a *scoring function* that measures how well that structure fits the data. The objective is to find the structure with the highest score. The difficulty arises when considering that the number of possible structures is superexponential ($2^{O(n^2)}$) to the number of variables[5]. As much as a scoring function can be defined, it is not clear how the optimal structure can be found. There are cases in which it is possible to find the optimal graph, however usually the problem is $\mathcal{NP}$-hard, and therefore heuristic search techniques must be implemented. Compared to the first approach, for which the failure of a single independence test risks compromising the entire structure, this one, having to optimize the entire structure instead, is less sensitive to individual failures. It also manages to better compare the benefit brought to the scoring function with the cost incurred when considering whether to add an edge. Clearly, the risk is to find a suboptimal solution.
When implementing score-based structure learning, it is important to choose the right scoring function. An alternative, as in parameter estimation, is the likelihood function. The idea is to find the model that makes the data at hand more likely. At this point one could try to maximize the likelihood of the entire model, that is, find the pair (K, θ_K). In other words, find the graph that achieves maximum likelihood, and thus makes the data as likely as possible, when assigning the parameters the values resulting from MLE. The problem with this choice is that maximum likelihood never prefers the simpler model over the more complex ones. This implies that this scoring

[5]The exposition of the concepts of the theory of complexity illustrated in this section is deferred to the Appendix.

function is inclined to choose complex models, whereby for example there are lots of edges in the face of very weak dependencies, and thus to overfit. This does not necessarily mean that maximum likelihood should be discarded by default, but that it should be implemented with caution, perhaps by providing complexity penalty mechanisms.
Another possible choice as a scoring function comes from a Bayesian approach. As done in parameter estimation, a priori distributions must be defined for the structure, $P(K)$, and on the parameters given the structure, $P(\theta_K \mid K)$. Applying the Bayes' rule, stated in Equation (6):

$$P(K \mid D) = \frac{P(D \mid K)P(K)}{P(D)}, \tag{28}$$

and, recalling that the denominator is a normalizing factor, the *Bayesian score* is defined as:

$$\text{score}_B(K : D) = logP(D \mid K) + logP(K). \tag{29}$$

Defining a prior distribution for the structure allows the choice of the scoring function to be directed, as it is possible to reward certain structures, for example more sparse, rather than others. However, it is the first term that is most relevant, and can be rewritten as:

$$P(D \mid K) = \int_{\Theta_K} P(D \mid \theta_K, K)P(\theta_K \mid K)d\theta_K, \tag{30}$$

where Θ_K is the set of the possible values of the parameters. This term is the marginal likelihood of the data given the structure. It is called marginal likelihood because it marginalizes out the unknown parameters of the Bayesian Network, considering all possible configurations of the parameters rather than a specific value. This is achieved by combining the first term of the integral, the likelihood of the data given the entire network, i.e. the pair (K, θ), with the second, the prior distributions over the parameters given the structure, effectively weighting each configuration of the parameters according to its prior probability. The difference with MLE is substantial, as the latter returns the maximum of the function, while the marginal returns the average based on the priors. This explanation illustrates why the Bayesian score does not overfit, unlike MLE, which in fact is too optimistic in its evaluation and returns parameters that are the best only if they are truly representative of the data in general, a rather rare circumstance. The Bayesian score, on the other hand, provides a set of parameters and a measure of how likely each is. Integrating $P(D \mid \theta_K)$ with respect Θ_K results in measuring the *expected likelihood*, weighted by all the possible values of θ_K. The estimate of the goodness of the model is thus more conservative. As a final consideration regarding the marginal likelihood, in the general case of estimating the structure of a Bayesian Network, it is usual to assume the independence parameter already illustrated in the previous section, for the same reasons.
The Bayesian score has the characteristic of preferring simpler structures, particularly compared with MLE. This does not preclude the possibility that complex structures cannot be estimated against data that require it. This scoring function balances the fit of the structure to the data with

complexity and tries to return the best trade-off. In this way overfitting is usually avoided. In case Dirichlet priors are used for parameters, then, for $J \to \infty$, so as the number of instances increases, it is possible to derive this approximation:

$$logP(D \mid K) = l(\hat{\theta}_K : D) - \frac{logJ}{2}\text{Dim}[K], \tag{31}$$

where Dim$[K]$ is the model dimension, i.e. the number of independent parameters in K. The score increases when the likelihood goes up, which measures how well the structure fit to the data, but decreases when the second term, that is proportional to the complexity of the model, specifically to the number of parameters and instances, rises. This shows how this scoring function tries to balance these two aspects and this approximation is called the *BIC score* (that stands for Bayesian Information Criterion).

For implementing the Bayesian approach for structure learning, as done in this analysis, one must therefore specify prior distributions for structure and parameters. Regarding P(K), it was already stated that, as useful as it may be in directing the scoring function toward some structures rather than others, it is less important. It is usual to assign a uniform prior. As for the prior on parameters, on the other hand, it is first useful to note that a desirable property is *score decomposability*, whereby the total score of a structure given the data is equal to the sum of the scores associated with all the variables given their parents. Intuitively, this is desirable because in the search for the best structure it is possible to operate locally, thus on individual nodes, without altering the score associated with the rest of the Network. For the score to be decomposable, another property, called *parameter modularity*, for which the prior of the CPD of a node depends only on the local structure, must be met. Not all parameter prior distributions satisfy this condition. This reduces the possible alternatives, which number nevertheless is still exponential. One of these is a Dirichlet$(\alpha, \ldots, \alpha)$, with a fixed value for each hyperparameter. If $\alpha = 1$, this prior is the already mentioned for parameter estimation K2 prior. Another possible choice already discussed is the BDe prior. In this case, the network structure associated to P$'$ is used only for specifying parameter priors, not for the structure search directly.

The last point to explore in this section is how the search for the best structure takes place in practice. Formally, given the data D, the scoring function, a set of possible structures $\mathcal{K}$ and the incorporated prior knowledge, the problem is to find the structure from $\mathcal{K}$ that maximizes the score. The choice of the scoring function is irrelevant, the important point is that it is decomposable. Finally, the *score equivalence* property applies: if K is I-equivalent to K$'$, then $score(K : D) = score(K' : D)$.

Once the variables have been selected, to build the structure one must decide which edges should be added and which should not. Suppose one wants to connect node X with node Y, following this direction, because it fits well to the data. This implies that the reciprocal edge, from Y to X, cannot be added, because it would result in a cycle. This simple case should not be particularly difficult to handle. However, the addition of this edge precludes connecting Y to Z and Z to X. Each individual addition thus has ramifications and implications beyond the individual node and the local structure.

Formally, defined an integer d and $\mathcal{K}_d = \left\{K : \forall i, |\operatorname{Pa}_{X_i}^K| \leq d\right\}$, i.e. is the set of graph for which each node can have at most d parents, it can be shown that, given a data set D and a decomposable score function, finding $K^* = \arg\max_{K \in \mathcal{K}_d} \text{score}(K : D)$ is $\mathcal{NP}$-hard for any $d \geq 2$. This shows that it is unlikely that there is an efficient algorithm for learning the best structure for any data set. This implies that it is generally needed to implement heuristic search algorithms, for which there is little guarantee of the goodness of the result. Faced with this consideration, it is possible to reformulate the problem by considering a search space for which the belonging graphs have a fixed limit of parents and the goal is to return one with a high score. The approach adopted is *local search*, for which the algorithm explores the search space without seeing all the elements, given that, as highlighted at the beginning of the section, its cardinality is superexponential to the number of variables.

Regarding the search space, it is useful to describe it as a graph, whose nodes are all the candidates in $\mathcal{K}$, which are connected on the basis of operators that applied allow them to move within it. The operator defines the operation to apply to the starting node that results in the candidate of the subsequent one. For the heuristic method of finding the solution to work, it is essential to define how the nodes of the graph are interconnected, that is, how the operators are defined. Thus, a trade-off arises: if in fact each operator allows to move from one candidate to a few others, it is possible to evaluate all these nodes, however it will take longer to arrive at the solution, compared to a graph in which from one node it is possible to move to several candidates, for which, however, it is more difficult to determine which is the right move. A good compromise involves arriving at an identical candidate to the starting one from each move except for local changes. Thus, there are three operators that define the connectivity of the search space graph:

- *edge addition*;
- *edge deletion*;
- *edge reversal*.

In addition, research is constrained by the maximum number of parents decided a priori, and only "legal" alternatives are allowed, so for example, the solution cannot involve cycles. The method looks for the best candidate within this space and moves between solutions by applying these operators.

The procedure of finding the best structure for a Bayesian Network is usually local. The one implemented in this analysis, called *greedy hill-climbing*, consists of these steps. A starting structure, which can be for example random or empty, is chosen, and the associated score is calculated. Neighboring nodes, which are obtained by applying the operators and assessing that they satisfy the constraints, are evaluated, and the modification that obtains the highest score is applied. If no change leads to a score increase, the resulting structure is the solution. From the point of view of computational cost, this way of proceeding is $O(n^2)$. This result is surely not very amusing. It is possible to improve it considering that most of the time an operator leads to a worse node. There

are algorithms that achieve lower costs, for example the first-ascent hill climbing, for which only a certain number of operators are sampled and the first resulting structure that gets a higher score is selected, immediately discarding the other operators. This procedure is faster and less costly, but also less accurate. Either way, with regard to the resulting structure, some considerations can be made. Any graph to which an operator is applied cannot return a higher score. So either a local maximum has been found, since all neighbors have a lower score, or a plateau has been reached. Based on this conclusion, it is possible to try alternative methods or to further improve the structure found; however, this is beyond the scope of this section.

3.4 Performing inference in a Bayesian Network

Performing inference in a Bayesian Network means, once the structure and parameters are learned, being able to derive the conditional probabilities of a certain event, for example being able to study the probability that a certain bank is under stress given that another is in that situation. The outcome is thus a *query* of the kind: $P(Bank_1 = \text{under stress} \mid Bank_2 = \text{under stress})$. These operations make it possible to study systemic risk in practice, since they quantify dependence among different banks. A Bayesian Network is able to return all these queries, but the inference problem in graphic models is $\mathcal{NP}$-hard and the same can even be stated for approximate inference[6]. This result therefore makes necessary to use alternative methods.

The *particle-based methods* are a class of methods that approximate the joint as a set of instantiations, called precisely particles. The idea is that it is possible to derive information from this representation of the general distribution. These methods are characterized by how particles are generated, ranging from deterministic techniques to sampling, and the definition of particles, which can be full particles, that is, complete instantiations on all variables, or collapsed particles, which instead involve assignment to only a subset of them. Formally, given a distribution $P(X)$, the objective is to estimate $P(Y = y)$, where Y is a subset of X and y is a possible assignment of Y. More generally, the goal is to estimate the expectation of a function $f(X)$ relative to P. To approximate this expectation, M particles are generated, the value of the function or its expectation relative to each generated function is estimated and the result is finally aggregated.

The approach adopted in this analysis is called *forward sampling*. It implements random sampling as the method of the generation of the particles. The outcome is thus M instantiations from the distribution $P(X)$. The sampling is implemented by following the topological order of the Network, so that for each node it is possible to draw according to the distribution conditioned to the values of the parents. For those without a parent, a value is drawn from the marginal.

[6] The exposition of the concepts of the theory of complexity illustrated in this section is deferred to the Appendix.

Thanks to basic convergence bounds, given the collection of the generated particles $D = \{\xi[1], \ldots, \xi[M]\}$:

$$\hat{E}_D(f) = \frac{1}{M} \sum_{m=1}^{M} f(\xi[m]), \tag{32}$$

whereas if the objective is to estimate $P(y)$:

$$\hat{P}_D(y) = \frac{1}{M} \sum_{m=1}^{M} \mathbf{1}\{y[m] = y\}, \tag{33}$$

where $\mathbf{1}$ is the indicator function that takes the value 1 if the condition in the argument is satisfied, 0 otherwise.

What has been highlighted demonstrates the simplicity with which probabilities can be estimated using these methods, especially when compared with the conclusion reached if one assumes to perform the exact or approximate inference. Important annotations must be added to what has been illustrated. Intuitively, the goodness of estimation of particle-based methods depends on the number of instantiations generated. However, it is not possible to give an estimate of how many are needed to get a good result. A trade-off as usual arises: forward sampling, like the other sampling methods for performing inference in Bayesian Networks, is surely easy to apply, but it is very difficult to ascertain if the obtained estimates are accurate.

This section concludes the chapter on the statistical model implemented in the analysis. What was illustrated is instrumental in understanding the choices made during the analysis. The next chapter discusses the results obtained.

Chapter 4

Results

4.1 General Overview

In this chapter the results of the analysis are presented. First, a Bayesian Network was estimated over the entire period from 2021-02-01 to 2023-01-31. The structure was estimated with the algorithm of greedy-hill climbing with BIC as scoring function. Next, parameters were estimated using the Bayesian method with the BDe prior. Once the whole Network was built, the forward sampling algorithm was applied. The number of particles generated was 10000. Each of these operations was accomplished with material provided by the Python library *BNLearn* (20).
Once the particles were generated, it was possible to perform inference to obtain the estimates of the conditional probabilities $P(i|j)$, i.e. for each bank j it was estimated the probability that every other bank i would have recorded a modified ϵ-drawup given that it had been recorded for j. In other words, the probability that CDS spreads referring to bank i had a significant growth, and thus the market perceived a situation of financial stress, given that the same was perceived for j. In practice, the calculation of probabilities was done by following what was done in [18]:

$$P(i \mid j) = \frac{n_{i,j}}{n_j}, \tag{34}$$

where $n_{i,j}$ is the number of samples in which for both banks a modified ϵ-drawup had been registered, and thus a *1* was present in the data, and n_j is the number for which a drawup had occurred only for the bank j. Another measure of systemic risk that was calculated was the expected number of banks under stress since bank j was under stress. For each sample, if for bank j a drawup was recorded, it was stored the number of other banks in a similar situation. In the end, the mean was computed over all the considered instantiations.
In addition, the same procedure was repeated but taking into account that a bank i could record a drawup even in the days following the stress of the conditioning bank j. This implied that for $P(i|j)$ it was also considered the case where bank i had 0.5 as its value in the computation, just as for

E[Number of Banks in Distress | j] the banks with that value were also counted as in distress given that j had a modified ϵ-drawup. In fact, estimates of the probability and number of banks under stress expected were higher. The idea behind this repetition was to compare the results taking into account or not stress propagation among banks.

Let this procedure be called *dtp* (day-to-period), while the previous *dtd* (day-to-day). Once the estimates were concluded, it was possible to analyze the probabilities for different banks, both in the role of i and j and for both procedures, and to rank the banks according to the conditional proportion of banks expected under stress, again for both dtd and dtp. This procedure was implemented only in the case of the estimation of the single Bayesian Network, whereas in the later sections dtp was not performed, for reason explained below.

Two Bayesian Networks were then estimated by dividing the dataset in half. Finally, 19 half-yearly Networks were estimated using the rolling window method, for which, at the end of each estimate over a period along the predetermined duration, the time window is moved so that the next half-year begins and ends one month later than the previous one. This method is useful for capturing fluctuations in the data, particularly when the process is non stationary (8). For the latter estimate, it was possible to analyze the time dynamics of systemic risk through the probabilities and proportions obtained from the networks, both at the aggregate and individual bank level. An additional analysis was conducted to visualize at what sub-period the risk was concentrated in the 2 years.

Finally, it is important to emphasize that it was never assumed that the hypotheses made in the Data Analysis part (chapter 2) should have been necessarily confirmed by the results. Different approaches were implemented to do a comprehensive analysis and to be able to state with some degree of confidence whether, through the available data and the implemented model, there were indications of any historical trends in systemic risk in the period under study and whether certain banks were more affected than others.

4.2 Estimating a single Bayesian Network on the whole period

Estimating a single Bayesian Network on the whole period implies discarding the time dimension, as it becomes meaningless. In Figure 8 the estimated structure for the Bayesian Network on the whole period can be visualized.

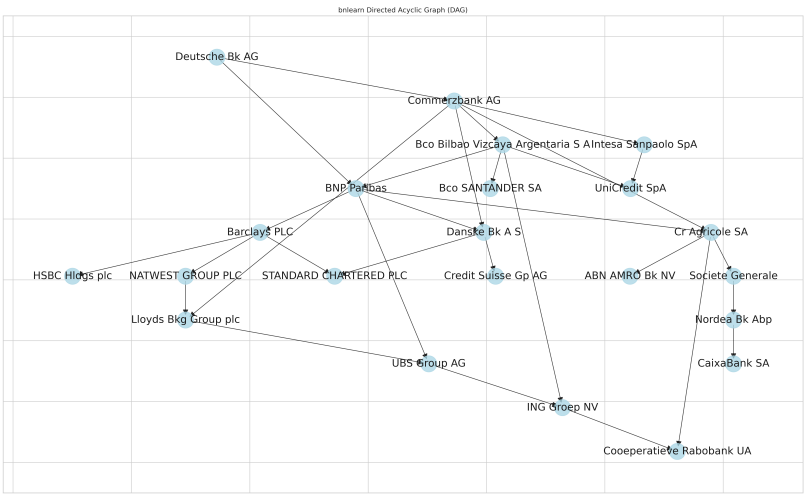


Figure 8: The estimated structure for the Bayesian Network on the whole period.

First, there seems to be a tendency for banks from the same nation to be linked. This could be a way to visualize how integrated the European banking system is beyond individual countries. Second, *Svenska Handelsbanken AB* is not part of the Network[1]. One possible interpretation of this phenomenon is that this bank in the period studied was independent of the others, that is, the stress of the other banks did not condition its situation and vice versa. This implication, which is probably the result of using BIC as a scoring function, is interesting because it allows to hypothesize that using this methodology one can derive only the critical entities for a period against a much larger set of variables. The model essentially could self-select only important banks to study systemic risk. Once the parameters have been estimated and inference performed, the probabilities and proportions obtained from the two different procedures can be analyzed.

[1]What was obtained was verified by discarding all the data about the Scandinavian bank and then re-estimating the structure of the Network. The outcome was the same.

Consider now the estimation of the mean of the conditional probabilities, starting from dtd, represented in Figure 9.

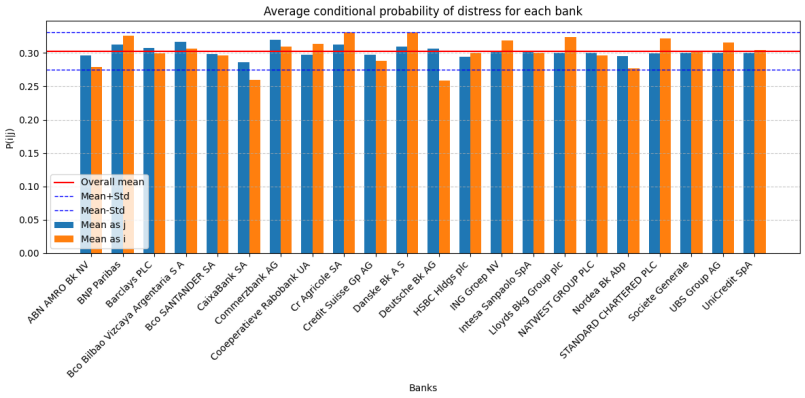


Figure 9: The estimated probabilities for a single Network with dtd procedure.

First, for each bank the probability estimate varies with respect to the average, around 0.3, whether it is the conditioned or the conditioning. This discrepancy can be explained by taking into account that there are banks more likely to influence others' stress than to be influenced, and vice versa. For example, *Deutsche Bk AG* is a bank for which its stress is more likely to lead other banks to go through turbulence than the other way around, probably also because it is the first bank in the structure, which is therefore not causally dependent on others. Instead, *Cr Agricole SA* and *Danske Bk A S* are the banks for which the estimated probability of going under stress given that another bank is under stress is higher on average. From the general point of view, it is interesting to note that the standard deviation is about 0.028, but unbundled by following the distinction in the plot it is about 0.008 when calculated by considering the probabilities for each bank in the role of j, while it is 0.02 in the role of i. In the second case there is more volatility, as can be guessed from the plot. Thus, for the first case, if all banks are considered individually to analyze the average probability that the others will go through a similar situation given one's stress, there is little difference. As much as there are banks with higher probabilities, the range is quite narrow, so on average each bank amplifies stress similarly. The other standard deviation shows which banks are more likely to go under stress when others go through such a situation, and thus in a sense are more susceptible to the cascade effect. These banks are interesting for studying systemic risk, as their stress is more likely to indicate general stress and ongoing contagion.

Analogously, Figure 10 represents the estimated probabilities with the dtp procedure.

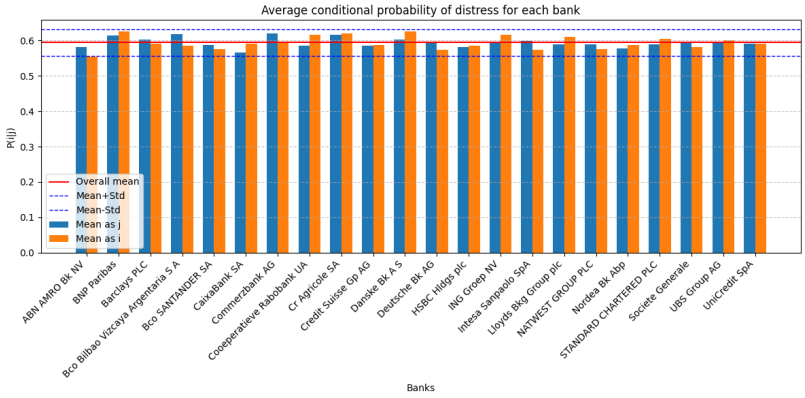


Figure 10: The estimated probabilities for a single Network with dtp procedure.

The same conclusions drawn earlier seem to apply here as well, however there is less volatility, which makes distinctions between banks and between probabilities referring to the same bank less pronounced and interesting.
Once the probabilities have been estimated, it is possible to compute the expected conditional proportion of banks in distress (see Figure 11).

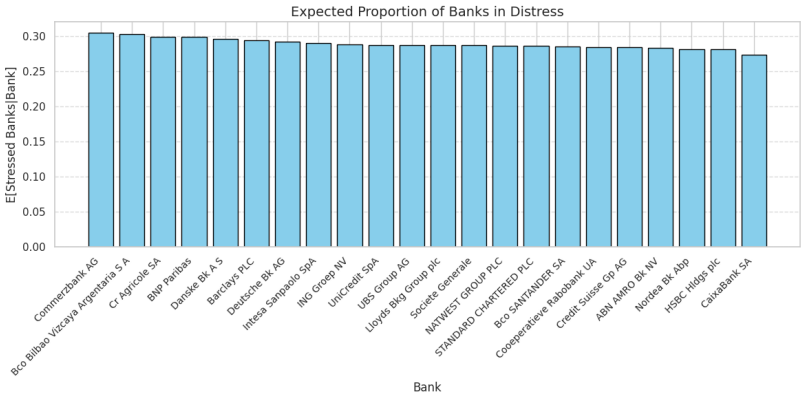


Figure 11: The estimated proportions for a single Network with dtd procedure.

What emerges is quite consistent with what was observed earlier. The variability is quite small and the ranking positions fairly reflect the probability estimates. It is noteworthy that the tendency found earlier, whereby banks from the same country are connected in the graph, does not imply that they have similar risk estimates, in addition to the fact that for *Credit Suisse Gp AG*, for which there had been a lot of stress in the data from a certain period onward (chapter 2.2), the estimates are quite small. This is probably also due to precisely the decoupling between the Swiss bank and the others from October 2022, because since the spreads of *Credit Suisse Gp AG* were very unstable but

the ones of the other banks were going back to lower values, the model did not detect dependence among the stresses.
Figure 12 reports the estimated proportions for the dtp procedure. The ranking is more or less the same and so also the conclusions.

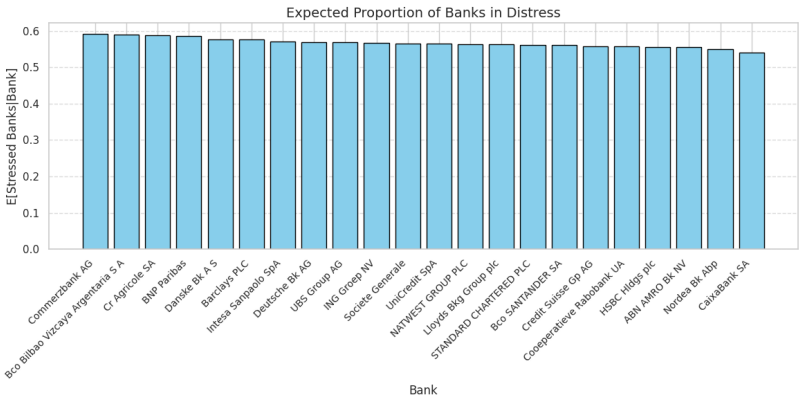
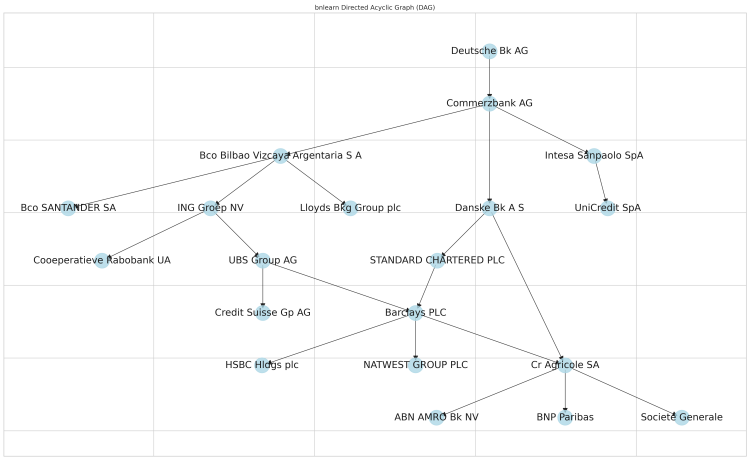


Figure 12: The estimated proportions for a single Network with dtp procedure.

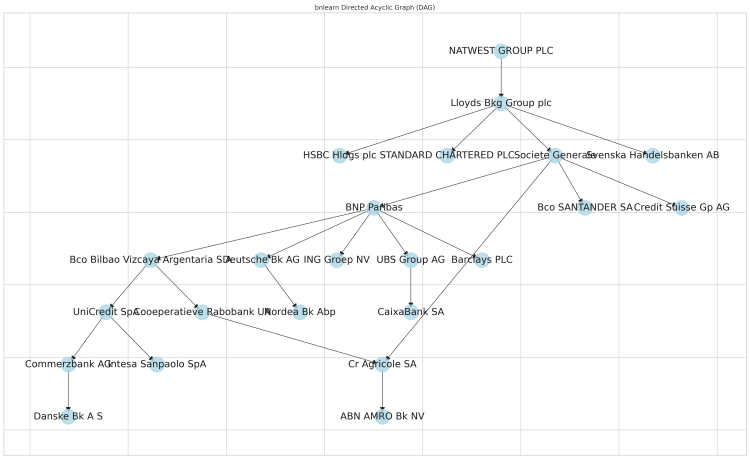
Estimating a single Network implies losing the possibility to study how the systemic risk changes during the considered period. It is plausible in fact that there have been some changes and to detect them a more granular method is needed. The results of these procedures are explained in the following sections, in which the results of the dtp procedure are not reported. The initial idea of considering also a small delay for stress propagation proved to be of little use, since results are just less volatile, and thus less interesting and meaningful, and the dtp procedure is even more complex than dtd, as an additional effort must be made to insert the 0.5 values and then to count them when estimating probabilities and proportions.

4.3 Estimating two yearly Bayesian Network

The first step in estimating two different models is to divide the dataset into the two periods one wants to compare. Here the idea was to compare the first year of the sample, from 2021-02-01 to 2022-01-31, during which spread values remained fairly stable, with the second, from 2022-02-01 to 2023-01-31, when instead there was a pronounced growth and generally strong instability. Thus, the results are expected to show a sharp increase in systemic risk from year one to year two. First, the two structures were estimated, as depicted in Figure 13.



(a) The estimated structure for the Bayesian Network on the first year.

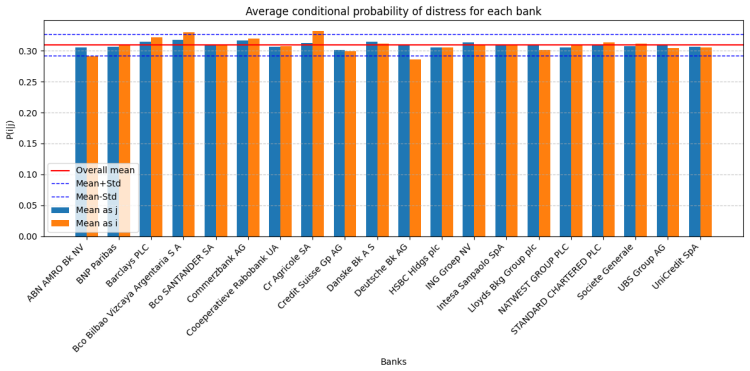


(b) The estimated structure for the Bayesian Network on the second year.

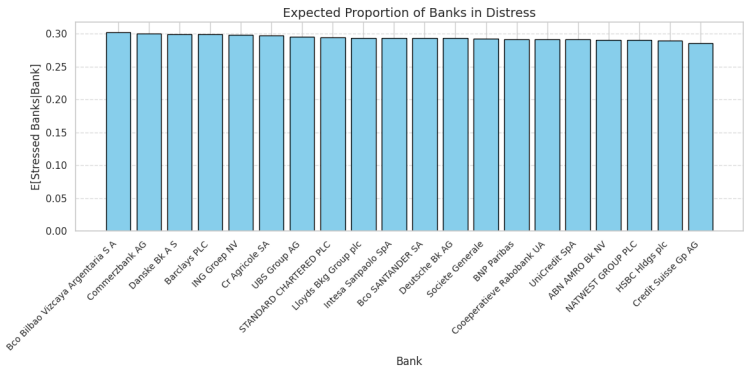
Figure 13: The two different structures estimated for the two different sub-periods.

In the first structure, *Deutsche Bk AG* is confirmed as the first node, and thus characterized by its own marginal distribution and causally conditional independent of the others. The tendency

of banks of the same country to be connected is also present. In general, there is some similarity with the estimated Network over the entire period. Interestingly, however, three banks are absent: *CaixaBank SA*, *Nordea Bk Abp* and again *Svenska Handelsbanken AB*. This growth relative to the individual Network is probably due to the fact that spread values in the first year were less volatile and thus fewer variables are needed to explain the data. This statement is also supported by the fact that, on the other hand, all banks are present in the estimated structure for the second year. In general, the topology is quite different, and there also seems to be less of the tendency regarding nationalities. Once the parameters have been estimated and inference made, it is possible to assess how much systemic risk has changed between the two periods according to the models.



(a) The estimated probabilities for the Bayesian Network on the first year.

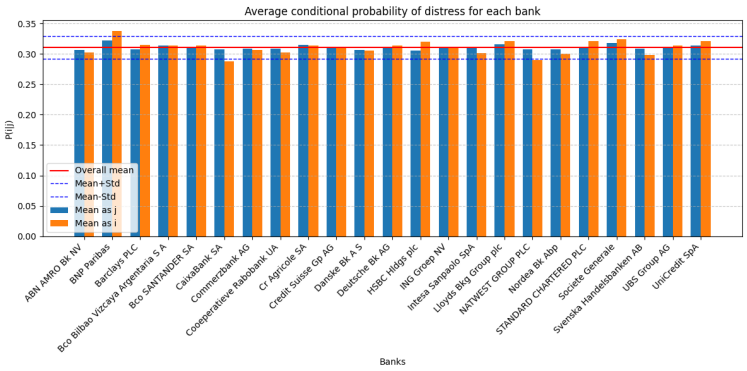


(b) The estimated proportions for the Bayesian Network on the first year.

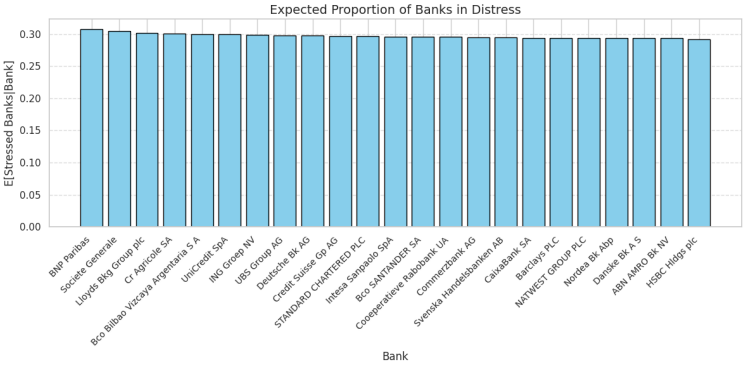
Figure 14: The two plots for analyzing the systemic risk in the first year.

In Figure 14, the results of the first year confirm what was already hypothesized when considering the structure, since there is some similarity with the results of the entire Network. However, the mean is higher (almost 0.31) and the standard deviation is smaller. It is also noticeable the growth of the probability of *Bco Bilbao Vizcaya Argentaria S A*, which in fact results in the bank with the

largest conditional number of banks expected to be under stress, while *Credit Suisse Gp AG* is on the opposite.



(a) The estimated probabilities for the Bayesian Network on the second year.



(b) The estimated proportions for the Bayesian Network on the second year.

Figure 15: The two plots for analyzing the systemic risk in the second year.

On the other hand, in Figure 15, the mean and standard deviation are essentially unchanged from the previous year. This seems counterintuitive, as instability was greater in the second year and thus one would expect systemic risk to have increased. The difference that emerges most clearly is the strong increase in the probability of *BNP Paribas*, also confirmed by the plot of proportions. Otherwise, few other conclusions can be drawn. An entirely different method was attempted at this point to try to investigate whether systemic risk actually changed over the two years, as intuitively the series plot seems to suggest (Figure 6). The idea was to enhance the time dimension even more through the rolling window method.

4.4 Estimating Bayesian Networks with the rolling window

The rolling window method for estimation was implemented considering six months, since it was considered the smaller possible amount of time to have a sufficient number of observation to return a robust result. Once this sub-period is selected from the general data set, the structure is estimated as usual. Considering that the total number of estimated Networks was 19, the graphical representation part was discarded, as it was difficult to synthesize the results for such a large amount of information. The excluded banks, however, were stored, as it was considered interesting to compare what was obtained with the Networks in the previous sections. Finally, the parameters were estimated and inference was performed as before. Once the semester estimate was completed, the window was shifted by one month, so the new sub-period began and ended one month later. This made it possible to verify changes in systemic risk in a granular manner, as the update of the interval for successive Networks was gradual, while still maintaining a sufficiently large amount of observations. The first information obtained once the entire estimate was completed concern the excluded banks for each semester. The three banks that repeatedly failed to be included in the structure were those already excluded in the Network of the first year: *Svenska Handelsbanken AB* 13 times out of 19, *CaixaBank SA* 10 times and *Nordea Bk Abp* 6 times. Clearly such high frequencies in a sense invalidate the results regarding these three entities, even in cases where they were included. Lastly, *HSBC Hldgs plc* and *Credit Suisse Gp AG* were excluded both once. Interestingly, considering the list reported in Appendix, there does not seem to be a strong correlation between the size of the bank and the times it was excluded. The importance of a bank, measured in this case by its size, thus does not imply that it is always crucial in the study of systemic risk, since it is possible to explain data concerning co-occurrence of stresses without including it. In addition to the frequency of exclusion, it was possible to analyze the semesters in which the structures were incomplete. For all sub-periods up to the one from 2022-01-01 to 2022-06-30 at least one bank was excluded. For the following four semesters, however, all variables were represented as nodes, and then in the last three again at least one bank discarded. The macroperiod for which all banks were included begins at 2022-02-01, the month in which the invasion of Ukraine took place, and ends with the semester May-October 2022. The Networks perfectly captured the time window in which the data show strong variability. It needed to be investigated, however, whether the risk estimates also turned out to be more relevant than those obtained in the rest of the sample, as it was hypothesized in the Data Analysis part (chapter 2).

The first part of the results analysis dealt with probability estimates. The idea was to partially replicate what was done in [16] with the *Joint Probability of Distress* measure. Once all the conditional probabilities were collected with the corresponding date, data were aggregated and the plot in Figure 16 was obtained.

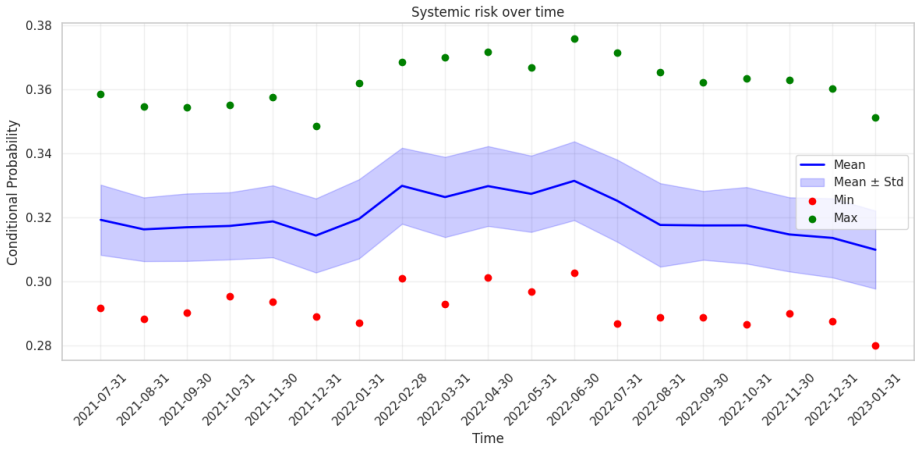
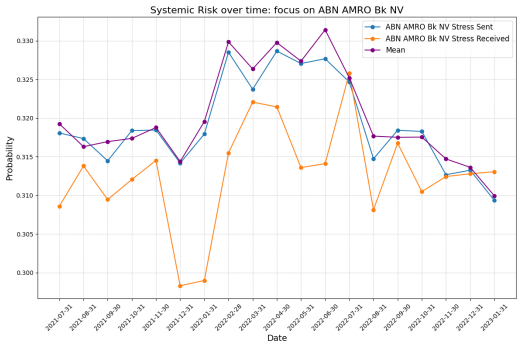
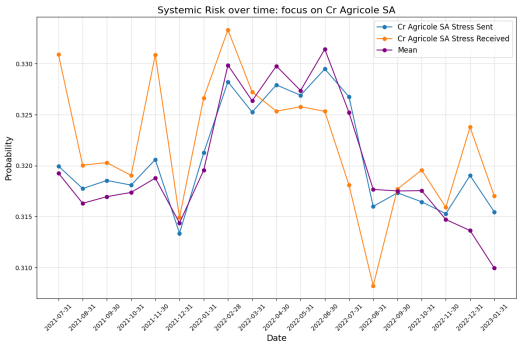


Figure 16: The evolution over time of the mean of the conditional probability $P(i|j)$.

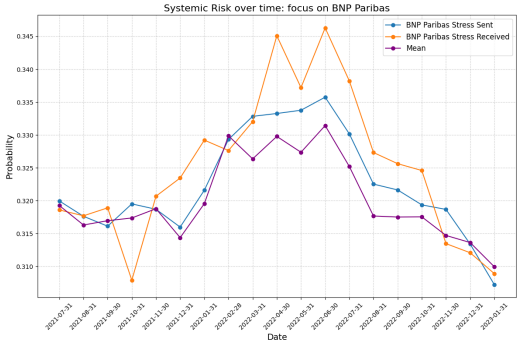
Figure 16 shows that from December 2021 the average probability of bank going under stress given that another was in that situation began to grow visibly and settled at a higher level until June 2022, when it then started to decrease. It is possible to speculate, given the coincidence in time, that the rise is due to the Russian-Ukrainian crisis, which has continued to produce instability effects on the European banking system for more than a quarter. Once what happened at the aggregate level was shown, it was possible to investigate the conditional probability evolution for individual banks as well. First, the three banks with an exclusion frequency well above 1 mentioned above were discarded because they were missing estimates for too many semesters. After that, for each bank, the probabilities were unbundled so that their evolution could be analyzed separately when the entity was the conditioning and when the conditioned. If one considers that when it is the conditioning it is assessed how much stress it sends to others and when it is the conditioned how much stress it receives, then it is possible to classify banks according to which for the most part is characterized by one behavior or the other. The idea was to partially replicate what was done by [15]. Finally, with the overall average shown in Figure 16, it was possible to calculate on average which banks deviated the most from the general trend and which were the most representative. In Figure 17 plots of four banks among others are reported: *ABN AMRO Bk NV* is the one that for most of the time sent stress and *Cr Agricole SA* is the one that for most of the semesters received it; *BNP Paribas* is the second less representative bank after *ABN AMRO Bk NV*, whereas *Standard Chartered Plc* is the most similar to the general mean. It is interesting to note some heterogeneity in the plots, for example between February and June 2022.



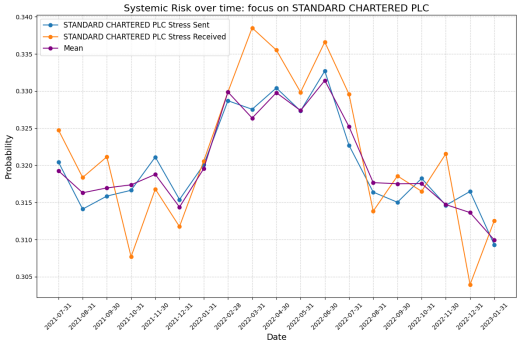
(a) The time evolution of the probabilities of *ABN AMRO Bk NV*.



(b) The time evolution of the probabilities of *Cr Agricole SA*.



(c) The time evolution of the probabilities of *BNP Paribas*.



(d) The time evolution of the probabilities of *Standard Chartered PlC*.

Figure 17: The plots of four banks with noteworthy results.

The latest analysis brought forward with probability estimation tried to retrieve a clearer picture of how much systemic risk had changed and to which banks it was principally attributable in the sample period. What has been done previously in fact has the merit of showing the general dynamics in the two years considered and then allows the situation of each bank to be put into context. It is also useful in assessing which banks are more representative of the general trend, which are more likely to destabilize others and which are more likely to be stressed by others. What is missing is perhaps a clearer indication of which bank is the most important for systemic risk, meaning the one most likely to be involved in conditional stress situations. To have this information available, the first percentile of the estimated probabilities was selected, thus the highest 1%. At that point for each date the relative frequency of these probabilities was calculated, so as to visualize in which period they were concentrated. In addition, for each probability, it was stored which bank held the role of conditioning in order to state which one should be attributed the greatest likelihood of generating stress in the system in the face of a strong growth in spreads. This made it possible to rank banks on the basis of which were responsible for the highest conditional probabilities.

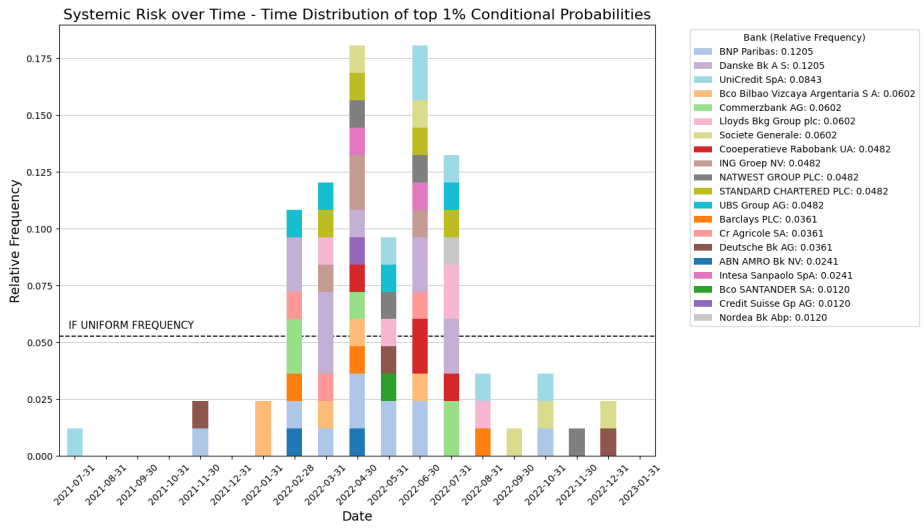


Figure 18: The concentration of systemic risk over time and the contribution of each bank.

Figure 18 shows that systemic risk was particularly concentrated in the period between February and July 2022[2]. The black line depicted highlights the expected height of the histogram bars if the top 1% probabilities had distributed evenly, and thus if systemic risk had remained more or less constant over the two-year period. In short, this plot confirms that systemic risk in the two years under consideration has been particularly concentrated since about the outbreak of war in Ukraine for about a little less than six months. As for the banks most responsible for the instability of the system, *BNP Paribas* and *Danske Bk A S* are tied for first place. It is noteworthy to mention that a similar result regarding the French bank was also found in [15]. Three banks, *CaixaBank SA*, *Svenska Handelsbanken AB* and *HSBC Hldgs plc*, do not even appear in this ranking, thus indicating that for no bank did the conditional probability of undergoing a stress given that these three experienced distress belonged in the top 1%. This supports the previous hypothesis that the importance of a bank, measured in this case by its size, does not necessarily imply that it is central to systemic risk. Given these results, it is interesting to notice the discrepancy between *HSBC Hldgs plc*, the largest bank in the sample, and the second one, *BNP Paribas*.
Similar analyses have been conducted regarding the expected number of banks in distress given a bank's stress, trying to partly replicate what was done in [16] with the *Expected Proportion in Distress*. Again, once the various proportions were estimated, the results were aggregated by date.

[2]The values are slightly approximated so the sum does not exactly make 1.

This made it possible to represent the average evolution over time (Figure 19).

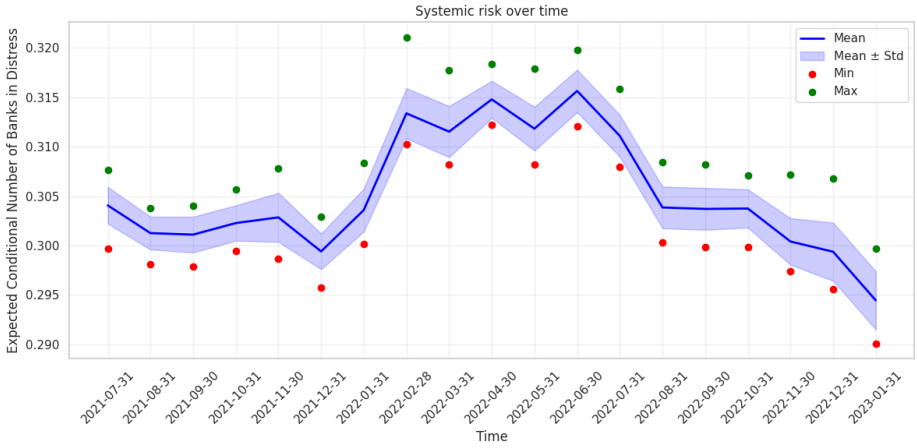
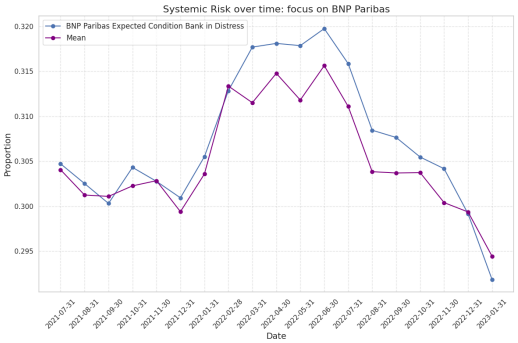
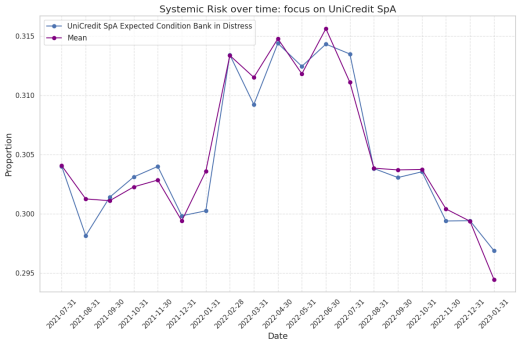


Figure 19: The evolution over time of the mean of the conditional expectation of banks in distress.

The plot is quite similar to the one in Figure 16, with the strong growth from December 2021 to February 2022, the settling to a higher level until June, and then the descent at the end, perhaps in this plot even more pronounced than in the previous. In short, the conclusions drawn before are also applicable to these results. In this case, however, the standard deviation is smaller, and this is noticeable when the data are represented at the individual bank level. In fact, the difference between different banks is less pronounced. It is anyway possible to state which bank is the most representative, *UniCredit SpA*, and which bank deviates most from the average, again *BNP Paribas* (see Figure 20).



(a) The time evolution of the proportions of *BNP Paribas*.



(b) The time evolution of the proportions of *UniCredit SpA*.

Figure 20: The plots of two banks with noteworthy results.

Using the rolling window method, it is possible to state that between 2021 and 2023 the systemic risk in the European banking sector did not remain constant. In particular, there has been a rise since the outbreak of war in Ukraine and a subsequent decline to the pre-crisis level. The banks most responsible for the amplification of risk were *BNP Paribas* and *Danske Bank A S*.

Chapter 5

Conclusions and future work

Two main conclusions can be drawn from the work done:

1. the systemic risk of the European banking sector has changed during the period 2021-2023, in particular has increased around the invasion of Ukraine by Russia and has remained stably higher for four to five months before returning to the previous level;

2. some banks were more responsible for the propagation of stress in the system, in particular *BNP Paribas* and *Danske Bank A S*.

Clearly, the results were influenced by the simplifications made upstream. First, it was decided to select banks based on their size, as this criterion was assumed to reflect the entity's contribution to systemic risk. Interestingly, the results partially disprove this hypothesis. *HSBC Hldgs plc*, the largest bank in the sample, played a marginal role in the evolution of systemic risk over the period, while the second one, *BNP Paribas*, is one of the banks that emerges as particularly relevant. If this phenomenon were to be confirmed by further studies, its causes could be investigated, which could prove useful for industry insiders. This aspect could be also investigated strongly to have better tools available to decide which banks should be given more attention. From this point of view, Bayesian Networks have the merit of "self-selecting" the most important variables. This phenomenon allows to study connections between banks or other entities simply by estimating the structure of the Network and comparing different topologies over time. A practical example that has been raised is to evaluate the integration of the European banking system.
Other simplifications made involved the values of some fields. Of greater significance might be the one concerning *Currency*, as the others are rather specific to CDS and therefore of less interest. The choice of selecting only the Euro was also supported by the fact that the study was about Europe. If it had been decided to consider the world banking system, this choice would hardly have been appropriate. These choices also tie in with the subsequent discarding of those banks whose data were too few or of dubious quality. First, it is unlikely that less restrictive choices on fields would have "cured" these issues. Second, the absence of some banks certainly affects the robustness of the

result negatively. It is likely, however, that the trend in detected systemic risk would have remained similar, as the exclusions are numerically less than the sample used, while it is not possible to state that those banks were not precisely the most important for systemic risk.
One aspect to be investigated is the detection of non-stationarity of the series (chapter 2.2). Usually in a TSA it is crucial to test this hypothesis (8) and, if unit roots emerge, to put some measures in place, as it is done for example in [16]. Stationarity is desirable because it allows what is detected in one sample to be inferred to the whole series. If, on the other hand, the statistical properties change significantly, what was discovered is unlikely to be applicable outside the context under investigation. As already raised, however, in [18] the issue is not addressed even if it is presumable that stationarity does not hold in that case either. Even in the present analysis this issue is not addressed, since it is not clear whether Bayesian Networks require data transformation, and in case of doubt it is preferred to maintain the approach adopted in the literature. The lack of clarity on this point is probably the main drawback of using Bayesian Networks for a TSA, at least from a methodological point of view. It is desirable that this topic can be deepened in a future work. In general, a certain opacity is found in the literature in the illustration of the available data and in the choices made. For example, there is a generic reference to spreads without clarifying which ones the author is referring to. Greater transparency is desirable to allow other authors to evaluate what has been done and to be able to build on the work of others.
Moving to the results obtained, it should first be highlighted that, for a study of the evolution of a phenomenon over time, the estimating of a single Network or in general of a limited number for non-overlapping time periods seems ineffective. In fact, the estimate of two annual Networks did not allow to detect any change in systemic risk, although this evidence subsequently emerged. Using a rolling window instead seems promising, although the advantage of the intuitiveness of the joint probability distribution representation is lost, as it is difficult to summarize the information returned by the plots. Another critical issue that emerged is the one related to the dtp procedure, which proved to be ineffective. Designing a method that allows for incorporating the delay in stress contagion between banks remains a primary issue and likely limits the estimates obtained.
The estimates obtained with the rolling window, from which the main conclusions drawn derive, nevertheless raise some questions. First of all, the estimated probabilities do not have any benchmark. It is therefore complicated to contextualize what has been achieved compared to the other crises that have been mentioned. It would be interesting to understand, for example, how much these probabilities have changed compared to the previous financial crisis. It is not clear though whether the implemented method is easily scalable, even with more powerful means. This makes the results difficult to relate to other contexts, because estimates are missing. This means, for example, that it is not known whether these probabilities indicate a very high risk in absolute terms, because there are no estimates for other periods. This implies that a certain dynamic of systemic risk has indeed been witnessed in the two-year period and that certain banks are more responsible than others for this evolution, but it is not possible to affirm anything in absolute terms, because no similar measure of risk is available for other crises. Certainly, the design of a systemic risk index would make it easier

to compare different periods and systems. Further research could be devoted to this aim.
Finally, with respect to what has been stated regarding the usefulness of implementing networks to study economic-financial phenomena (chapter 1.3), it is interesting to note that in this case the loss of resilience prior to the tipping point was not found. Looking at Figures 16 and 19 until January 2022, it is difficult to see any growing weakness in the system. One reason could be that the previous time window is too short. Another one is that there was no phenomenon of this type because the crisis was not structural, but due to a sudden event that caused instability, which was then quickly reabsorbed. Of course, it may also be that this method is not the appropriate one, or that a higher granularity was needed to detect the loss of resilience. This consideration is supported by the growth witnessed between December 2021 and January 2022. It could be interesting to delve deeper into this topic to provide further useful elements to evaluate the usefulness of these tools in economic-financial analysis.
In conclusion, the results obtained provide a clear indication of what happened in the period analyzed, both to the general dynamics of systemic risk and to the role of the various banks. Some simplifications have been made upstream of these conclusions, which open up different scenarios for future work. In general, an attempt has been made to provide further elements in the ongoing search for models capable of predicting future periods of instability in the economic-financial system and their effects. Future work will try to clarify whether Bayesian Networks can be among these and, if so, with which data and methodologies should be adopted.

Appendix

The complete list of selected banks

Here is the list of the 30 largest European banks by assets reported in the *Europe's 50 largest banks by assets, 2022* (19) and the possible reason why the bank is not present in the results obtained from the Bayesian Networks estimation:

1. *HSBC Holdings PLC*;
2. *BNP Paribas SA*;
3. *Crédit Agricole Group*;
4. *Barclays PLC*;
5. *Banco Santander SA*;
6. *Groupe BPCE*: discarded due to the limited data available during the period of analysis;
7. *Société Générale SA*;
8. *Deutsche Bank AG*;
9. *Intesa Sanpaolo SpA*;
10. *Lloyds Banking Group PLC*;
11. *Crédit Mutuel Group*: discarded for the behavior shown in the time series plot;
12. *UBS Group AG*;
13. *ING Groep NV*;
14. *NatWest Group PLC*;
15. *UniCredit SpA*;
16. *La Banque Postale SA*: not present in the data when filtering the banks;

17. *Credit Suisse Group AG*;
18. *Standard Chartered PLC*;
19. *Banco Bilbao Vizcaya Argentaria SA*;
20. *CaixaBank SA*;
21. *Rabobank*;
22. *DZ Bank AG*: discarded for the behavior shown in the time series plot;
23. *Nordea Bank Abp*;
24. *Danske Bank A/S*;
25. *Sberbank of Russia*: not included in the analysis because it is Russian;
26. *Commerzbank AG*;
27. *ABN AMRO Bank NV*;
28. *KBC Group NV*: discarded due to the limited data available during the period of analysis;
29. *Nationwide Building Society*: discarded for the behavior shown in the time series plot;
30. *Svenska Handelsbanken AB*;

Notes on Complexity Theory

When evaluating an algorithm, it is important to take into account its *computational cost*, i.e., the time it takes to complete the calculation and the amount of space it requires (11). It is expressed in general terms and according to the size of the problem it is to solve. It is usually denoted by the notation $O()$, which indicates its *asymptotic complexity*. The study of this topic is called *complexity theory*. Once this cost is derived, based on the function stated within the parenthesis, the *running time* of an algorithm is defined. For example, if a problem has a computational cost $O(n)$, it means that it is expected that an algorithm can solve it in a time linearly proportional to its size. Among the different possibilities, it is usual to distinguish between *polynomial time* and *exponential time*. In fact, in the former case it is considered feasible to implement that algorithm to solve the problem; in the latter case it is not, in particular when n is big. This distinction is so important that it is the criterion for defining *equivalence classes of problems*. Given that two problems belonging to the same class are equivalent, if one algorithm is shown to be efficient in solving one problem, i.e. in polynomial time, it is possible to solve the other in the same way. The problem arises from the fact that for many problems there do not seem to be algorithms that can solve them in polynomial time. One class of equivalent problems is called $\mathcal{P}$, for which there is a deterministic algorithm that solves

them in polynomial time. Another class is $\mathcal{NP}$, for which on the contrary a solution can be verified in polynomial time. In other words, a nondeterministic algorithm, which means that it is designed to firstly formulates a guess and that takes it as the candidate, can verify the latter deterministically in polynomial time. Since deterministic algorithms are a special case of nondeterministic algorithms, it is possible to state that $\mathcal{P} \subseteq \mathcal{NP}$. The reverse, which would mean that, for every problem for which it is only possible to verify a solution in polynomial time, it is also possible to design an algorithm that solves it deterministically in polynomial time, is one of the biggest open problems in computer science. The point is to understand whether, even in the case of finding a polynomial solution to one problem, then one could claim to have found a polynomial solution to all the problems in the class $\mathcal{NP}$.

The $\mathcal{NP}$-complete problems are the most difficult one in the class, since the search for an efficient solution has so far been unsuccessful. Stating that a problem belongs to this class means that finding a polynomial solution for it would imply showing that there exists an efficient algorithm also for all the others in the class, which seems very unlikely. In practice, the notion of $\mathcal{NP}$-hard problems is used, which are problems that are at least as difficult as the $\mathcal{NP}$-complete. They might not even belong to $\mathcal{NP}$, so not always a candidate can be verified efficiently, but any problem of the latter class can be reduced to it in polynomial time. The classic way to demonstrate that a problem probably does not have an efficient solution is thus to show that it is $\mathcal{NP}$-hard. This is done by finding a reduction from some known problem that belongs to this class to the problem under investigation. If this happens, the implicit statement is that, if there is an efficient solution for the problem under study, this would imply that there would be a polynomial solution for all $\mathcal{NP}$-complete problems, and so that $\mathcal{P} = \mathcal{NP}$. This is believed to be unlikely.

Bibliography

[1] I. Anagnostou, S. Sourabh, and D. Kandhai. Incorporating contagion in portfolio credit risk models using network theory. *Complexity*, 2018:1–13, 2018.

[2] Albert-László Barabási. *Network Science*. Cambridge University Press, 2016. Available at `https://www.networksciencebook.com/`.

[3] Stefano Battiston, J. Doyne Farmer, Andreas Flache, Diego Garlaschelli, Andrew G. Haldane, Hans Heesterbeek, Cars Hommes, Carlo Jaeger, Robert May, and Marten Scheffer. Complexity theory and financial regulation. *Science*, 351(6275):818–819, 2016.

[4] Matteo Bernardi and Leopoldo Catania. Switching generalized autoregressive score copula models with application to systemic risk. *Journal of Applied Econometrics*, 34(1):43–65, 2019.

[5] Giovanni Bulfone, Roberto Casarin, and Francesco Ravazzolo. Corporate cds spreads from the eurozone crisis to covid-19 pandemic: a bayesian markov switching model. RCEA Working Paper 21-09, Rimini Centre for Economic Analysis, 2021.

[6] Wikipedia contributors. Exploratory data analysis. `https://en.wikipedia.org/wiki/Exploratory_data_analysis`, 2024. Accessed on December 29, 2024.

[7] Wikipedia contributors. List of largest banks. `https://en.wikipedia.org/wiki/List_of_largest_banks`, 2024. Accessed on December 29, 2024.

[8] Rob J. Hyndman and George Athanasopoulos. *Forecasting: principles and practice*. OTexts, Melbourne, Australia, 3rd edition, 2021. Accessed: 2025-01-03.

[9] A. Johansen. Characterization of large price variations in financial markets. *Physica A: Statistical Mechanics and its Applications*, 324:157–166, 2003.

[10] R. Kaushik and S. Battiston. Credit default swaps drawup networks: Too interconnected to be stable? *PLoS ONE*, 8(7):e61815, 2013.

[11] Daphne Koller and Nir Friedman. *Probabilistic Graphical Models: Principles and Techniques*. MIT Press, Cambridge, MA, 2009.

[12] Markit Group Limited. Data dictionary & automating report downloads for markit cds pricing, 2017. Confidential. Copyright © 2017, Markit Group Limited.

[13] IHS Markit Ltd. Cds indices primer, 2021. Confidential. Copyright © 2021, IHS Markit Ltd.

[14] IHS Markit Ltd. Composite cds basic pricing, 2022. Confidential. Copyright © 2022, IHS Markit Ltd.

[15] George Moratis and Plutarchos Sakellaris. Measuring the systemic importance of banks. *Journal of Financial Stability*, 54:100878, 2021.

[16] Dong Hwan Oh and Andrew J. Patton. Time-varying systemic risk: Evidence from a dynamic copula model of cds spreads. *Journal of Business and Economic Statistics*, 36(2):181–195, 2018.

[17] Fabio Saracco et al. Detecting early signs of the 2007–2008 crisis in the world trade. *Scientific Reports*, 6:30286, 2016.

[18] Snehashish Sourabh, Markus Hofer, and D. Kandhai. Quantifying systemic risk using bayesian networks. Technical Report 3525739, SSRN, 2020.

[19] SP Global Market Intelligence. Europe's 50 largest banks by assets, 2022. Accessed: 2024-12-29.

[20] Erdogan Taskesen. Learning Bayesian Networks with the bnlearn Python Package., January 2020.

[21] Wikipedia contributors. Credit Suisse — Wikipedia. `https://en.wikipedia.org/wiki/Credit_Suisse`, 2024. Accessed: 2025-01-02.