UNIVERSITY OF PANNONIA

DOCTORAL THESIS

Gravity Models and Machine Learning Approaches for Understanding Corporate Investment Flows and Framework Programme Collaborations in Europe

DOI:10.18136/PE.2025.947

Author:
Ferenc KIRÁLY

Supervisor:
Prof. Dr. habil. Zsolt Tibor KOSZTYÁN

A thesis submitted in fulfillment of the requirements for the degree of Doctor of Philosophy

in the

Doctoral School in Management Sciences and Business Administration
Department of Quantitative Methods

July 17, 2025

Declaration of Authorship

I, Ferenc KIRÁLY, declare that this thesis titled, "Gravity Models and Machine Learning Approaches for Understanding Corporate Investment Flows and Framework Programme Collaborations in Europe" and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Signed: _____

Date: _____

Gravity Models and Machine Learning Approaches for Understanding Corporate Investment Flows and Framework Programme Collaborations in Europe

Thesis for obtaining a Ph.D. degree in the **Doctoral School in Management Sciences and Business Administration** of the University of Pannonia

in the field of **Social Sciences**
in the subject of **Management and Business Studies**

Written by: Ferenc KIRÁLY
Supervisor: Prof. Dr. habil. Zsolt Tibor KOSZTYÁN

Propose acceptance (yes/no) .....
Supervisor

As a reviewer, I propose acceptance of the thesis:

Name of reviewer: ..... yes / no

.....
(Reviewer)

Name of reviewer: ..... yes / no

.....
(Reviewer)

The Ph.D. candidate has achieved .....% at the public discussion.

Veszprém, ...
(Chairman of the Committee)

The grade of the Ph.D. Diploma .....(.....%)

Veszprém, ...
(Chairman of UDHC)

UNIVERSITY OF PANNONIA

Abstract

Doctoral School in Management Sciences and Business Administration
Department of Quantitative Methods

Doctor of Philosophy

Gravity Models and Machine Learning Approaches for Understanding Corporate Investment Flows and Framework Programme Collaborations in Europe

by Ferenc KIRÁLY

European economic integration and research collaboration patterns require systematic analysis to understand underlying network formation mechanisms and predict future developments. This study was designed to model European ownership networks and Horizon 2020 collaboration networks, with the objectives of improving link prediction accuracy, identifying key influential factors in network formation, and detecting outlier communities that deviate from predicted collaboration patterns.

A comprehensive research database was constructed by integrating four major data sources: CORDIS (Community Research and Development Information Service), Amadeus/Orbis (corporate ownership databases), PATSTAT (European Patent Office database), and Eurostat (European statistical data). For ownership network analysis, 1,620,340 companies with verified ownership relationships across 1,435 NUTS 3 regions were examined between 2010-2018. A gravity-based null model was developed and applied to predict spatial structures of corporate ownership networks. For collaboration network analysis, data from over 25,000 H2020 projects involving more than 150,000 participants were analyzed using multiple machine learning approaches including Random Forest (RF), Support Vector Machines (SVM), XGBoost, Linear Discriminant Analysis (LDA), Quadratic Discriminant Analysis (QDA), and logistic regression. Bayesian optimization was employed for hyperparameter tuning, and feature importance was assessed using the Boruta algorithm.

The gravity-based model demonstrated superior performance compared to existing Newman-Girvan and Expert models in estimating global network properties and centrality metrics. For collaboration prediction, non-generic machine learning methods achieved higher accuracy than generic approaches, with Random Forest yielding the best performance (F1 score of 1.302585). Outlier collaboration communities were detected, predominantly comprising organizations from EU core countries, indicating stronger-than-predicted collaboration patterns within established research networks.

The integrated modeling approach successfully improved link prediction accuracy for both ownership and collaboration networks while revealing systematic patterns in European network formation. The gravity-driven modularity measure enabled identification of economically coherent communities and their temporal evolution. The findings provide actionable insights for policymakers seeking to optimize collaboration efficiency in Framework Programmes and support strategic decision making in corporate investment flows across European territories.

Acknowledgements

I would like to express my deepest gratitude to my professor and supervisor, Prof. Dr. Zsolt Tibor Kosztyán, whose invaluable guidance, encouragement, and expertise have been the cornerstone of my PhD journey. Your unwavering support and insightful feedback challenged me to grow both as a scholar and as an individual. I am sincerely thankful for your patience, inspiration and the many opportunities you provided me to develop my research skills. This achievement would not have been possible without your mentorship. Thank you for believing in me and for being an exceptional advisor throughout my studies.

I am profoundly grateful to my family for their unwavering support and encouragement throughout my PhD journey. To my late father, whose memory continues to inspire and guide me - thank you for the values and determination you instilled in me. Although you are no longer here, your influence has been a constant source of strength.

To my mother, thank you for your unconditional love, patience and faith in my abilities; Your constant encouragement has been my foundation.

To my brother, András, your encouragement and understanding, especially during challenging times, have meant more to me than words can express. Thank you for always believing in me and for being a source of motivation and joy.

Most importantly, to my wife, Viktória, thank you for your endless love, patience, and sacrifices. Your faith in me and your constant support have been my greatest strength. I am truly grateful for your companionship, understanding and sharing this journey with me. This achievement would not have been possible without you by my side.

Finally, I appreciate all the valuable feedback and discussions from peer reviewers, proofreaders and experts.

Professional acknowledgements

PREPARED WITH THE PROFESSIONAL SUPPORT OF THE DOCTORAL STUDENT SCHOLARSHIP PROGRAM OF THE CO-OPERATIVE DOCTORAL PROGRAM OF THE MINISTRY OF INNOVATION AND TECHNOLOGY FINANCED FROM THE NATIONAL RESEARCH, DEVELOPMENT AND INNOVATION FUND.

Project no. KDP-11-3/PALY-2021 has been implemented with support provided by the Ministry of Culture and Innovation of Hungary from the National Research, Development and Innovation Fund, financed under the KDP-2020 funding scheme

NATIONAL RESEARCH, DEVELOPMENT AND INNOVATION OFFICE
HUNGARY

PROGRAM
FINANCED FROM
THE NRDI FUND

Contents

List of Figures

List of Tables

Acronyms

AI	Artificial Intelligence. 10
BA-model	Barabási-Albert (Scale-free) Network model. 7, 8
BC	Betweenness centrality. 57, 81
CF	Cash flow. 38, 76
CI	Corruption index. 32, 37, 38
CN	Common Neighbor Method. 23
COM	Company Ownership Matrix. 39
CON	Company Ownership Network. 45, 60
CR	Current ratio. 38, 75, 76
csv	Comma Separated Values file. 32, 33
DA	Discriminant analysis. 50
DC	Degree of centrality. xiii, 57, 81–83
EC	Earned contribution. 37, 38, 76, 79
EFTA	European Free Trade Association. 31
EIC	Economic-investment communities. 5, 44–46, 91
EN	Number of employees. 38, 76
ER-model	Erdős-Rényi Random Graph Model. 7, 8, 18–20
ERGM	Exponential Random Graph Model. 19, 20
EVC	Eigenvector centrality. 57
FAs	Fixed Assets. 38, 75
FP7	7^{th} Framework Programme. 37, 38, 79
FPs	Framework Programmes. 1, 4, 89, 91
GCN	graph convolutional network. 4, 24
GDP	Gross Domestic Product. 20, 32, 38
GEN	Gravity-based Economic Null model. 5, 41, 42, 44, 48, 60, 66, 68, 72, 85, 89, 91, 95
GIS	Geographic Information Systems. 9
GNN	Graph Neural Network. 4, 24
H2020	Horizon 2020 - EU research and innovation funding programme 2014-2020. 1, 3, 5, 17, 26–28, 49, 85, 89, 91, 95

ICCR	Institutional Cumulative Cooperation Ratio. 22
IDF	Institutional Document Frequency. 22
KGE	Knowledge Graph Embeddings. 4, 24
LDA	Linear discriminant analysis. xv, 50, 51, 54, 75–77, 79–81, 86
LogR	Logistic regression. xv, 50, 53, 54, 75–77, 79–81, 86
NGO	Non-Governmental Organization. 16
NLP	Natural Language Processing. 22
OR	Operation revenue. 38, 74, 75
P/L	Profit and loss. 38
PI	Number of patents. 37
PLF	P/L before tax. 38, 74–76
PLF	P/L for period. 38, 75, 76
PM	Profit margin (%). 38, 75
Pov	Persons at risk of Poverty or social exclusion. 32, 37
PPP	Purchasing Power Parity. 34
QDA	Quadratic discriminant analysis. xv, 51, 76, 79–81, 86
RB	ROE using P/L before tax (%). 38, 75
RBF	Radial basis function. 51, 52
RCB	ROCE using P/L before tax (%). 38, 75, 76
RF	Random Forest. xiii, 52–54, 56, 77–82, 86
ROCE	Return On Capital Employed. 38
ROE	Return Of Equity. 38
SH	Shareholder Funds. 38, 75
SME	Small and Medium sized Enterprise. 27
SNA	Social Network Analysis. 8, 15
SP	Shortest Path. 47
SR	Solvency Ratio (asset based) (%). 38, 75
SVM	Support Vector Machine. 51, 52, 54, 56, 79–81, 86
TA	Total Assets. 38, 74, 75
VIF	Variance Inflation Factor. 42
XGBoost	Extreme Gradient Boosting. 53, 54, 56, 79–81, 86

Chapter 1

Introduction

1.1 Research goals

One of the ultimate objectives of the study is to understand and model the European ownership network to derive valuable information on how investment flows within the European Union and what factors influence the formation of this network. The goal is to improve the link prediction together with improving the derived network parameters for which a gravity-based null model is proposed and identify the main investment communities together with identifying the time stability of them. Achieving this goal provides a deeper understanding of the factors that play a role in the formation, and this information can be beneficial to decision makers and government representatives to support regional and national economic decision-making processes.

The other main goal is to gain a deeper understanding of the collaboration network of the Horizon 2020 - EU research and innovation funding programme 2014-2020 (H2020). The aim is to provide an accurate model that can predict the collaboration links between firms based on measurable numeric information and identify the influential parameters that play a role in the formation of the network. The construction of an accurate link prediction model is regarded advantageous for decision makers seeking to forecast future consortia within the Framework Programme. Identifying key influential factors enables facilitation and support of future Framework Programmes (FPs) collaborations by decision makers. Furthermore, outlier collaboration communities-characterized by links that are either stronger or weaker than predicted-are also identified in this study, providing relevant insights for the management and regulation of collaboration formation across the Framework Programme. These considerations are consistent with the strategic directions of the European Union, which emphasize the improvement of collaboration between stakeholders at both the country and the organizational levels (Novitzky et al., 2020).

1.2 Motivation of the thesis

The investigation of ownership networks serves multiple critical motivations that intersect with economic, regulatory, and societal factors. First, analysis of ownership networks sheds light on hierarchical structures within corporations and their interlinkages, elucidating the control dynamics over assets and decision-making processes. Understanding these structures is fundamental for various stakeholders, including policymakers, as it informs regulatory frameworks aimed at enhancing transparency and accountability within corporate governance (Lidth Jeude et al., 2019; Takes et al., 2018).

Moreover, ownership networks reveal complex interdependencies among corporations that can lead to monopolistic behavior and anti-competitive practices, which have significant implications for market efficiency and economic equity. Identifying key players within these networks helps regulators understand potential risks, such as orchestrated collusion or financial manipulation, providing an avenue for intervention (Mizuno et al., 2020; Nakamoto et al., 2019; Villamil et al., 2024).

From a methodological perspective, several studies have employed network science techniques to explore ownership structures, leading to the development of innovative frameworks that capture the layered nature of corporate hierarchies. Such frameworks improve our ability to analyze the complexity embedded in ownership relationships, particularly in transnational settings where ownership stakes can be obscure and dispersed (Babić et al., 2019; Rungi et al., 2017). This complexity is crucial to understanding how power is concentrated within a small number of corporations, thus shaping competitive dynamics on a global scale (Vitali et al., 2011).

The benefits of investigating ownership networks are extensive, impacting not only academic discourse but also practical implications for governance and financial regulations. Through improved transparency and accountability mechanisms, stakeholders can foster environments that are conducive to fair competition and sustainable economic growth (Kosztyán et al., 2022b).

The investigation of collaboration networks in research is driven by several motivations, mainly centered on enhancing knowledge transfer, fostering innovation, understanding social dynamics within scientific communities, and maximizing research impact. These networks facilitate the interaction between researchers, influencing their collaboration choices and overall productivity.

The critical motivation for analyzing these networks stems from the desire to understand and potentially rectify social and epistemic inequalities in science. For example, Li et al. (2022) propose that a comprehensive understanding of collaboration factors can illuminate the inequalities that persist within scientific communities, thus supporting research and development efforts. Collaboration networks provide insights into how researchers associate with each other, which is crucial to address disparities in access to resources and opportunities.

In addition, the structural configurations of collaboration networks can also influence the dissemination of innovation. Katerndahl (2011) discusses how external collaborators play an important role in network dynamics, allowing a wider dissemination of innovations. This aligns with the findings of Vanni et al. (2014), who note that international collaboration networks can generate more precise epidemiological estimates, illustrating the importance of cross-border collaboration in global health contexts.

Collaboration networks serve as mechanisms for knowledge transfer, particularly between developed and developing nations. Long et al. (2015) assert that a shared understanding of goals and network structures among collaborators is a key indicator of successful collaborative efforts that improve collective results. This contextualizes the need for effective frameworks that structure collaborations, especially in multidisciplinary settings.

One of the primary motivations behind exploring collaboration networks is to enhance R&D productivity and innovation performance. For example, Laufs et al. (2024) examine the role of prior knowledge and existing collaborations in accelerating R&D performance during urgent contexts such as the development of the COVID-19 vaccine, indicating that extensive networks can facilitate the faster mobilization of resources and knowledge essential for rapid innovation. Similarly, Kong et al. (2017) demonstrate that collaborative approaches can alleviate challenges in

product R&D processes by allowing firms to leverage complementary skills and capabilities, ultimately enhancing innovation. These findings underscore the critical role collaboration plays in enhancing R&D efficiency.

Collaboration networks also serve as instruments for improving the performance of regional innovation. Roesler and Broekel (2017) show how universities act as central hubs in interregional R&D collaborations, linking local firms to broader knowledge networks, thus allowing increased resource allocation and collaborative innovation activities that contribute to regional economic development. This dynamic underscores the ability of such networks to generate synergies that may not be possible in isolated efforts.

1.3 Personal motivation

As an employee (Project Manager) of an international automotive supplier company with a headquarter in Germany, I was always interested when and more specifically why a company is establishing a new subsidiary in a different country, what are the main factors indicating this action. This question is not only interesting for me as an employee of a multi-national company but also as a Hungarian citizen, as for our country the attractiveness for foreign companies carrying investments into our homeland is one of the essential sources of the economic growth of the nation. So, the work I did during the study of the ownership network of EU companies was very interesting to find answers to the questions I had.

The collaboration network investigation was also an aspect of my daily life, as I was always curious about the factors influencing whether two (or more) companies are working together on one project to be more successful together than separately. During my job, day by day I work together with colleagues from different companies with whom our aim is common, but I was always interested in what could have been the reason why these particular companies are working together. The investigation of the H2020 collaboration network gave me some insight into the possible reasons, which were very interesting although the investigation was a huge task, but due to its complexity, I really enjoyed the work and the result of the work is satisfactory.

1.4 Research gaps

Considering the investigation of corporate ownership networks in the recent literature, several gaps can be identified. One of the main ones which was experienced during the literature review is that complete research was not performed until the date of publication of the current work (based on the best knowledge of mine), which would have investigated the ownership network of the European companies as a whole, integrating different data sources. Based on this finding, the current work tries to fill this gap by investigating the entire available corporate ownership network in the European Union.

In this area several methodological difficulties were also identified (such as temporal changes in the network, layered structure of the network, difficulty to determine centralities, etc.) which are mentioned in Section 2.3 and also referred to in the same section where the current study aims to address and answer these challenges.

Despite of the advances in link prediction methodologies reviewed in Chapter 2, prediction accuracy remains a persistent challenge in collaboration network analysis (Wang et al., 2015a). Two primary factors contributing to suboptimal accuracy are

identified: (1) limitations inherent to the models and (2) constraints associated with the datasets.

From a model-based perspective, the computational complexity of advanced methods such as Graph Neural Networks (GNNs), Knowledge Graph Embeddings (KGE), and graph convolutional networks (GCNs) poses significant scalability issues. Although these models demonstrate promise in link prediction tasks, their training and inference complexity escalate prohibitively with network size, creating bottlenecks for large-scale applications. Furthermore, the capacity of GNNs to capture long-range dependencies and intricate patterns in complex networks remains an active area of investigation. Furthermore, neural networks lack robust mechanisms to interpret the importance of characteristics, which is critical for strategic decision making. These observations are consistent with Chen et al. (2021), where the prediction precision between multiple algorithms plateaued near 0.75. The absence of high-accuracy prediction models hinders effective strategic planning at the organizational and consortium levels.

From the database side, incomplete or fragmented datasets contribute to inaccuracies in the prediction. No prior study has systematically integrated comprehensive databases such as ORBIS, CORDIS, PATSTAT, and EUROSTAT, resulting in omitted variables and relationships that degrade model performance.

The combined effect of these limitations makes reliable predictions for EU framework programme collaborations challenging.

Existing studies prioritize overall prediction performance while neglecting outlier detection-instances where observed collaborations deviate significantly (stronger or weaker) from predictions. Without analyzing such outliers, region or actor-specific insights cannot be derived, precluding targeted policy interventions in EU FPs. This impedes the ability of the EU to optimize collaboration efficiency.

1.5 Research questions

The research questions are formulated to provide a more focused, precise and researchable framework in contrast to the broader research objectives. In the dissertation, these research questions are addressed through a literature review and the development of a theoretical background. Based on this foundation, a research assumption is elaborated in subsequent chapters. Through the implementation of the research addressing the main points of the thesis, the assumptions are confirmed or rejected, and the corresponding research theses are articulated.

Taking into account the above issues and their relevance, the current study seeks to answer the following research questions.

> **RQ1:** Can the proposed gravity-based economic null model improve link prediction and network coefficient estimation, identify stable Emergent Innovation Communities, and provide insights into their spatial and temporal dynamics?
>
> **RQ2:** To what extent do administrative borders influence investment flows, and how do these effects change when controlling for geographical distance?
>
> **RQ3:** Can the proposed model, applied to a comprehensive dataset, enhance our understanding and predictive accuracy of organizational collaboration and community structures in Framework Programmes beyond current benchmarks?

In Section 7 a summary table is provided (Table 7.1) which collects the Research Questions, Assumptions and Theses for and overview.

For answering **RQ1** the literature review section provides some insight into the current challenges in the area of investigation. Within that chapter and later on during the introduction of the methods employed within the study, it will be revealed that several methodological improvements are needed to be able to provide deep insights for the question. The Gravity-based Economic Null model (GEN) model has been introduced that provides better link prediction than the other ones mentioned in this work, but the main benefit of this approach is the use of a purely economic model to predict the network, which highlights its importance.

Related to the **RQ2** the formation of investment flows within corporate ownership networks should be analyzed. To be able to address this challenge together with the aspect of the temporal viewpoint, so-called Economic-investment communities (EIC)s have to be identified and the structure of them must be analyzed for which modularity investigation is performed. Moreover, to investigate the time difference in such community formation, a multilayer approach is applied where the different layers represent the different years.

Looking at the **RQ3** context, in this work it is discussed and shown that the applied machine learning technologies -both generic and non-generic ones- are beneficial in collaboration network investigation, at least with respect to H2020 collaboration within Europe. The current status of the link prediction area is also introduced in a detailed way based on the literature (refer to Section 2.7) and also mentions the main challenges that are addressed in this study. It is also revealed that the different tools and methods are advantageous in different ways and applying these heterogeneous methods as a model, the prediction accuracy can be improved, and also the important forming factors can be identified. The usage of the generic methods are helpful in the analysis of corporate variables with which it can be revealed that organizations engaging in collaborative activities generally exhibit more favorable characteristics. The non-generic or black-box methods have the benefit to help to identify the important features and parameters which have effect on the formation of the network.

It is common related to all Research questions that the benefit of a comprehensive database is non-questionable. It is revealed in this work that the involved heterogeneous data sources is beneficial as most of the data originated from different databases has statistically significant explanation power in several aspects, therefore, the creation of the research database can also be considered as a mentionable result.

1.6 Structure of the thesis

The dissertation is organized as follows. After a brief introduction of the work in Chapter 1, in Chapter 2 the literature is reviewed with the related works and findings and based on them characterizes the research assumptions. In Chapter 3 the employed data with exact sources are introduced, which is followed by the explanation of the methods used. Chapter 4 explains the results of the work, while Chapter 5 discusses them and also defines the research theses. Chapter 6 shows the threats to the validity. Chapter 7 summarizes and concludes the dissertation and Chapter 8 considers the limitations and also proposes future research based on the current work.

Chapter 2

Literature review

2.1 Main milestones of Network Science

Network science is recognized as an interdisciplinary domain dedicated to the analysis of complex networks and their inherent properties. Over recent decades, several pivotal milestones have directed the progression of this field. An appreciation of these milestones is essential, as they establish the foundational context for contemporary investigations in network science.

Among the earliest and most influential models is the Erdős-Rényi Random Graph Model (ER-model), which was formulated between 1959 and 1968 by Pál Erdős and Alfréd Rényi (Erdős and Rényi, 1959, 1960, 1961a,b, 1963, 1966a,b, 1968). In this framework, a random graph is defined by a fixed number of nodes (N) and a probability (p) that determines the probability of an edge being present between any pair of nodes. The simplicity of the ER-model has allowed the derivation of fundamental results in relation to network connectivity and phase transitions. However, the ER-model does not account for several characteristics observed in empirical networks, such as the power-law degree distributions that typify systems such as the World Wide Web and social networks (Albert and Barabási, 2002; Glos et al., 2021).

Small-world problem was first mentioned by Milgram (1967), and small-world network was first formally described by Watts and Strogatz in 1998 Watts and Strogatz (1998) - although the mathematical effects were published much earlier in 1978 by de Sola Pool and Kochen (1978) -, are a class of graphs that exhibit high clustering (like regular lattices) and short average path lengths (like random networks). This unique structure, often termed the "small-world phenomenon," captures the idea that any two nodes in the network are connected by a surprisingly small number of steps, popularized by the "six degrees of separation" concept. The Watts-Strogatz model demonstrates how introducing a few random connections ("shortcuts") into a regular lattice dramatically reduces the diameter of the network while preserving local clustering. This property has profound implications for understanding systems ranging from social interactions to neural connectivity, as it balances efficiency and robustness.

Small-world topology has been observed in diverse real-world systems such as social networks inherently form small world structures, facilitating tight-knit communities and global information diffusion (Newman, 2003).

In contrast to the ER-model, the Barabási-Albert (Scale-free) Network model (BA-model), introduced by Réka Albert and Albert-László Barabási in 1999 Barabási and Albert (1999a), marked a significant advance by demonstrating that many real-world networks exhibit power-law degree distributions. The BA-model incorporates two principal mechanisms: network growth and preferential attachment, the latter signifying that new nodes are more likely to connect to those with already high degrees (Bollobás et al., 2011; Pedarsani and Grossglauser, 2011). This process

results in scale-free networks, where a minority of nodes (hubs) possess a disproportionately high number of connections, while the majority have relatively few. The BA-model's significance extends beyond its descriptive accuracy; it has also informed the understanding of the robustness and susceptibility of the network to random failures or targeted attacks (Albert and Barabási, 2002; Glos et al., 2021).

Throughout the evolution of network science, these foundational models have been integrated into broader analytical frameworks to better capture the complexity of diverse networked systems. Developments such as community detection algorithms, models of epidemic propagation, and the study of dynamic networks have extended the applicability of the ER-model and BA-model (Farzaneh and Coon, 2022; Rinaldo et al., 2013). These advancements have incorporated elements such as node heterogeneity, temporal evolution, and spatial constraints, underscoring the necessity for continual refinement and adaptation of theoretical models to reflect the multifaceted nature of real-world networks (Farzaneh and Coon, 2022; Međedović, 2020).

The dynamic interplay between theoretical modeling and practical application in network science has yielded substantial impacts across disciplines including sociology, biology, computer science, and epidemiology. Ongoing research continues to investigate the influence of network structure on processes such as disease transmission, information spread, and systemic resilience (Daudin et al., 2007). Furthermore, analytical methodologies derived from these theoretical models have facilitated the classification, visualization and interpretation of social, biochemical, and economic networks (Farzaneh and Coon, 2022; Međedović, 2020).

Social Network Analysis (SNA) serves as a visual methodology for characterizing network structures and nodal relationships by translating spatial systems into quantitative relational data (Ye et al., 2022). During the past two decades, SNA has gained prominence as a critical tool in regional science and economic geography (Hui et al., 2020). Its applications encompass investigations into urban and economic agglomeration patterns Liu et al. (2018), Searle et al. (2018), and Van Meeteren et al. (2016), alongside analyses of innovation networks, knowledge diffusion Abonyi et al. (2020), Czvetkó et al. (2021), Dahesh et al. (2020), Morrison (2008), Sebestyén and Varga (2013), and Weidenfeld et al. (2021), trade links Bhattacharya et al. (2008) and Mao and Cheng (2019), and tourism interactions (Asero et al., 2016; D'Agata et al., 2013; Liu et al., 2012a; Mou et al., 2020; Seok et al., 2021). In contrast, systematic modeling and analysis of network dynamics in international business research remains underdeveloped (Kurt and Kurt, 2020).

In summary, understanding the key developments in network science - particularly the ER-model and BA-model models - enhances the ability to analyze and interpret complex systems. These models have substantially shaped the theoretical landscape and have provided a solid foundation for ongoing innovation and interdisciplinary research within the field.

2.2 Economic networks

The evolution of economic networks forms a complex picture that combines historical contexts, theoretical frameworks, and practical implementations. To clearly articulate this evolution, it is essential to examine various facets, including the historical significance of trade, the theoretical underpinnings provided by game theory and economic modeling, as well as the implications of new technologies and regulatory frameworks on network structures.

The historical context of economic networks can be traced back to medieval Europe, where long-distance trade was pivotal to early economic growth. Institutions developed alongside merchant networks facilitated impersonal trade, which was instrumental in fostering economic development in Western Europe. According to Kallioinen (2020) research, the evolution of these networks corresponds to the establishment of legal frameworks that promoted individualism between corporations, allowing a rise in trade activity and, subsequently, the development of economic networks that continue to influence modern economies today.

In tracing the history of economic network research, scholars have identified several key transitions influenced by broader socio-economic transformations. Early economic thought rarely considered the relational dynamics that facilitate transactions. Instead, it emphasized market efficiency and individual decisions. This began to change with the advent of theories that recognized the intrinsic connectivity among economic agents. For example, Granovetter (2005) work on the social structure of economic outcomes demonstrated how embedded social networks profoundly influence hiring practices, pricing mechanisms, and innovation trajectories, ultimately redefining notions of market dynamics.

As these networks evolved, the application of economic modeling solidified its significance in understanding network interactions. Berry and Johari (2011) highlight that knowledge of game theory is imperative for engineers and economists alike, as it encapsulates the dynamics of network innovation and adoption. Within this scope, economic modeling allows for the analysis of interactions between users and providers in terms of incentives and behaviors, reinforcing the idea that networks are not merely technological constructs but are also deeply embedded in economic interactions (Walrand, 2008). Furthermore, economic regulation serves as a fundamental element in optimizing these networks, especially in industries that exhibit natural monopolies. Economists have long debated the efficacy of regulatory frameworks in ensuring competitive practices, as regulatory intervention is crucial to achieving allocative and productive efficiency in contexts involving monopolistic power (Uukkivi and Koppel, 2020).

The interplay of urban development and economic networks also has historical roots. The structure of cities has been heavily influenced by economic geography, which explains the spatial evolution of industries and the competitive dynamics that shape them. The work of Frenken and Boschma (2007) illustrates how evolutionary economic geography provides a perspective on how industrial dynamics emerge and influence urban growth, allowing for the analysis of how networks of competing firms coexist and interact within urban spaces through localized economies. The role of visualization in economic networks has been enhanced by advances in data analytics that allow real-time insights into the relationships that govern economic interactions. As seen in studies of the portrayal of urban networks by economic geography, the application of Geographic Information Systems (GIS) and spatial analysis has revealed how urban configurations can dynamically influence economic outcomes across diverse contexts (Batty, 2008). Such spatially explicit analyses yield insights into the relational dynamics that govern urban economic structures, reinforcing the symbiotic relationship between geography and economic performance. Similarly, analysis of corporate networks reveals the importance of understanding the nodal relationships between firms that span different geographies, as demonstrated Wall and Knaap (2011), who emphasize the hierarchical structures within global corporations that drive economic concentrations in specific urban centers.

The continuous blurring of boundaries between disciplines, particularly economics and sociology, has fostered a richer understanding of the role of networks.

Insights from network topology and graph theory have been applied to understand the structural properties of economic interactions and cluster formations within markets. For example, the sectoral differentiation examined by Djauhari and Gan (2016) employed Minimum Spanning Tree approaches to provide insights about the historical clustering of economic sectors, contributing to our understanding of intersectoral relationships.

With the rise of computational capacities, empirical investigations have started to layer complex mathematical and statistical methodologies that model interactions among economic entities. One notable methodological innovation is the application of machine learning techniques, such as neural networks, which have gained traction for economic predictions and optimizations, particularly in forecast modeling contexts (Liang et al., 2024). This computational advancement underscores a conceptual evolution in which traditional theories merge with data-driven methodologies, leading to improved forecast accuracy and more nuanced economic analyses.

The adaptive nature of economic networks has been significantly shaped by advancements in technology. The literature on Artificial Intelligence (AI) in economics indicates that as economic models evolve, so do the methodologies used to analyze networks (Bickley et al., 2022). The integration of AI systems into economic analysis allows for a more nuanced understanding of market behavior and network interactions, thus enhancing predictive capabilities and informing decision-making processes. This evolution is crucial in an era where data analytics becomes a cornerstone of economic modeling, particularly within networked environments.

At the same time, applying social network analysis reveals distinct patterns in urban and economic interactions. The critical role of various cities within broader economic systems has been brought to light through statistical networks that assess urban positionality and economic linkages. Sigler and Martinus (2016) examination underscores the importance of urban networks in driving economic activities, challenging previously held notions of hierarchical urban influence in favor of a more comprehensive understanding of cities as interconnected economic actors. The connectivity of cities through firm interactions reflects broader socio-economic trends that have historical precedents dating back to the early trade networks.

Another dimension of the evolution of economic networks can be attributed to the regulatory frameworks that have emerged over time. The systematic approach to regulating network industries showcases how legislative and institutional frameworks adapt in response to the unique challenges posed by network structures. Effective regulation is crucial to understanding the interaction between local contexts and regulatory instruments in fields such as telecommunications (Uukkivi et al., 2014).

As our understanding of economic networks continues to evolve, it becomes clear that contemporary models must incorporate the complexities of globalization, socio-political contexts, and technological advancements. From immigration's impact on trade networks to the infrastructural developments in port cities, the interconnectedness of local and global economies necessitates nuanced models that can account for diverse influences on economic structures. Walton-Roberts (2011) critique on Indo-Canadian transnational networks emphasizes how geographic and historical contexts can uniquely influence trade dynamics, suggesting that future economic network research must embrace this multifaceted approach.

Moreover, the application of urban scaling laws — highlighted in literature by authors such as Lobo et al. (2013) - positions cities as dynamic systems governed by similar structural rules despite their unique circumstances. This conceptualization parallels findings linking urbanization processes in different contexts, providing a

coherent framework to understand the interactions unfolding within economic networks.

In summary, the history and evolution of economic networks reveal a rich interplay of historical advancements, theoretical frameworks, and the inexorable march of technology. By analyzing early trade practices, applying economic models that factor in game theory, and recognizing the differing influences of urban and global dynamics, we can gain insight into the profound and intricate nature of these networks. This comprehensive approach underscores the importance of historical contexts and illustrates the need for adaptive, innovative frameworks in understanding and leveraging economic networks in today's complex world. The study of economic networks is inherently interdisciplinary, bridging insights from economics, sociology, urban studies, and emerging technologies.

2.3 Corporate ownership network analysis

In this section, several references from the literature are introduced that underscore the need and importance of investigating and researching corporate ownership networks. This part is relevant for one of the main goals of the current work and aims to show why the topic is important and relevant to research.

The exploration of ownership networks within corporate governance has gained significant traction in recent years, especially as scholars attempt to discern the intricate relationships between ownership structures and various corporate performance metrics. Ownership structures profoundly influence corporate governance practices and, in turn, the financial performance of firms. This discourse requires examining several aspects, including ownership concentration, stakeholder dynamics, and governance disclosures, all of which are illuminated through recent academic advancements. The importance of the field investigation is also highlighted by Pehrsson (2016) as the establishment of a new subsidiary firm is a formation of investment that is beneficial both on the source and the target side.

Ownership concentration has been shown to have a considerable impact on earnings informativeness. Specifically, Sharifi and Jafari (2016) found that concentrated ownership is related to low earnings information, as ownership concentration prevents leakage of proprietary information about firm rent seeking activities, which were prevalent and profitable in the selected firms in Asia. This lack of transparency can exacerbate agency problems, where the interests of minority shareholders may conflict with those of controlling shareholders.

Furthermore, the control exercised by certain owners significantly shapes governance structures, which, in turn, affects business ethics and corporate behavior. Zattoni (2011) discusses the importance of flexible corporate law that adapts ownership rights and governance structures according to the evolving characteristics of a firm. This flexibility enables firms to establish suitable governance frameworks that align with the interests of various stakeholders, allowing for adjustments as the company evolves.

The intersection of corporate governance and ownership is also illustrated by the nature and degree of voluntary disclosures made by companies. Ntim et al. (2012) showed that government ownership is positively correlated with the degree of voluntary disclosure of corporate governance among South African firms, indicating that ownership types influence transparency practices. This finding aligns with the notion that greater ownership diversity can enhance the scrutiny and monitoring of managerial actions, potentially leading to more robust governance outcomes.

In addition to government and institutional ownership, the distinction between ownership types can lead to different corporate strategies and risk profiles. Liu et al. (2024) examined the impacts of government ownership on the environmental preferences of private firms within emerging markets, concluding that state ownership contributes positively to environmentally responsible corporate practices. This finding implies that government participation in ownership structures can extend beyond financial oversight and contribute positively to corporate social responsibility outcomes.

As the study of ownership networks continues to evolve, scholars have increasingly recognized the challenges associated with indirect ownership and the multilayered structures that often underlie corporate networks. Romei et al. (2015) outlined the complexity involved in computing direct and indirect ownership relationships, emphasizing the importance of identifying corporate group structures controlled by parent shareholders. Understanding these intricacies is essential for evaluating the efficacy of governance mechanisms and identifying possible points of failure or conflict within corporate networks.

In addition, regional studies have highlighted the significant effects of corporate governance on financial performance in various contexts. For example, research by Darko et al. (2016) demonstrates that ownership structures greatly influence corporate outcomes, particularly within the banking sector in Ghana. The malleability of corporate governance practices in response to ownership types suggests that strategic adjustments could be made to align governance with performance.

At a meta-level, the interplay between ownership structures and corporate governance mechanisms can impact financial distress risk. Research by Kim (2019) suggests that concentrated ownership can serve as a mitigation factor against corporate bankruptcy, particularly in contexts with weaker institutional frameworks. This highlights the nuanced role of ownership concentration in providing stability, although at the potential expense of governance quality.

Ultimately, the dynamic nature of ownership networks requires ongoing investigation to uncover their complex relationships and influences on governance. As ownership structures continue to transform, driven by factors such as globalization and technological advancement, understanding these changes and their implications for corporate governance has never been more critical. The challenges and opportunities that arise from ownership networks will shape the future of corporate governance practices across various sectors around the world.

Several methodological challenges are also identified in the literature in the context of investigation and research on ownership networks.

Research on corporate ownership networks, particularly within the context of the European Union (EU), faces several methodological challenges that hinder clear understanding and comparative analysis of ownership structures and their implications on corporate governance, performance, and regulations. This complexity is amplified by the diverse ownership patterns exhibited in different jurisdictions, creating nuanced layers of interdependencies between corporations.

One of the primary methodological challenges in corporate ownership network research is the issue of data availability and quality. Many studies have highlighted the incomplete and often non-representative datasets that researchers must rely on when constructing ownership networks. Mizuno et al. (2023) highlight that the flow of corporate control is contingent on the robustness of ownership data available, and incomplete representation issues can lead to misleading conclusions about the actual influence wielded by certain corporate entities within the network. Garcia-Bernardo et al. (2017) further discuss the implications of these incomplete data, suggesting

that future research should focus not only on comprehensive ownership data but also on relationships and transactions between entities to provide a more rounded understanding of the ownership network.

Another critical challenge is the complexity of ownership structures themselves, particularly the prevalence of layered ownership and cross-holdings among corporations. This creates significant difficulties when it comes to establishing the true controlling entities within a network. The work of Vitali and Battiston (2014) and Vitali et al. (2011) emphasizes the concept of indirect ownership control and how intrinsic complexities can mask the visibility of true ownership. Multilayered structures often complicate the analysis of ownership dynamics, making it difficult for researchers to discern direct influences and control mechanisms between entities.

The problem of network centrality and its measurement is another methodological bottleneck. Studies often struggle with accurately determining the centrality of nodes within ownership networks, thereby affecting the interpretation of influence and power dynamics. This issue is highlighted by Vitali and Battiston (2013, 2014), who discuss how centrality measures can be contingent on the specific configurations of network structures and the underlying assumptions about ownership rights and control. Misinterpretation of centrality can drastically alter the conclusions drawn about corporate influence in economic and regulatory contexts.

The constantly evolving nature of corporate ownership - particularly given the rapidly changing regulatory frameworks and market dynamics in the EU - adds another layer of complexity for researchers. The research landscape is marked by ongoing changes in the laws and regulations that govern corporate ownership and control, necessitating continuous adaptation and reevaluation of methodologies (Vitali et al., 2011). This is crucial in order to keep pace with emerging trends, such as the shift toward more transparent ownership structures required by increasing regulatory scrutiny.

This challenge related to data availability is addressed in the dissertation Sections 3.1 and 3.2.1 to establish a comprehensive data set that contains most of the available data and with this database to be able to increase the reliability of the investigation.

The layered structure of corporate ownership related challenge is addressed also in the current work as the direct ownerships are considered as this information is available in the used Amadeus database. More details are mentioned in Sections 3.1.2, 3.1.3, and 3.2.1.

In this work, centrality identification, measurement, and usage related challenges are also considered, since several different types of centralities were calculated and used as referred to in Section 3.3.1.

The evolving nature of the ownership connections, especially the temporal differences are handled in the dissertation with applying multilayer networks as representation of the different years of ownership. Further details are discussed in Section 3.3.1.

In conclusion, the integration of ownership structures and corporate governance practices is of paramount importance to ensure organizational effectiveness and accountability. As research progresses, the emphasis on ownership types, their evolution, and the resulting implications for governance will provide vital insights into fostering sustainable corporate practices amidst an ever-changing business landscape.

2.4 Collaboration networks analysis

Collaboration network analysis plays a key role in understanding the intricate dynamics of partnerships in various fields, serving not only as a pathway to knowledge sharing, but also as a mechanism to address complex global challenges. The nature of these collaborative networks presents both methodological and practical challenges that are often underestimated. Recent literature highlights this multifaceted exploration that encompasses theoretical frameworks, practical applications, and the impediments associated with building effective collaborative networks.

The importance of collaboration network analysis in the economic field, particularly among companies and other economic entities, lies in its profound implications for innovation, competitive advantage, and sustainability. Firms increasingly recognize that they do not operate in isolation; Their performance and ability to innovate are significantly influenced by their relationships with other entities. This network of interactions, characterized by collaborative efforts between companies, universities, research institutions, and suppliers, plays a critical role in improving innovation and competitive performance.

Collaboration networks facilitate the exchange of knowledge and resources, that leads to significant advantages in innovative outputs. For example, Spender et al. (2017) found that ongoing alliances with various entities, including competitors and research institutions, can enhance radical innovation. This notion is supported by Nejad et al. (2013), who emphasize that companies should engage with a wide array of actors within the technological innovation system to maximize their innovative potential. Thus, a broader network perspective is essential for firms aiming to capitalize on diverse resources and expertise available within their collaborative landscapes.

Furthermore, the formation of these collaborative networks is crucial in responding to the dynamic nature of market demands. In the context of industrial production processes, Pinto Leão and Silva (2021) highlight that digital transformation enhances overall value chain integration and promotes distinctive competencies through collaborative ventures, ultimately strengthening firms' competitive advantages. Such transformations indicate that, in an increasingly digitized economy, the ability to collaborate effectively requires an adaptive approach that integrates diverse capabilities across the network.

In the digital age, the interaction between collaboration networks and technological innovation has taken on new dimensions. Zou and Xi (2024) assert that the structural characteristics of knowledge-sharing networks significantly influence technological innovation. This perspective underscores the evolving nature of inter-firm relationships and the necessity of adapting to the interconnected economic landscape.

Inter-firm collaboration, particularly in emerging markets, demonstrates variable outcomes depending on external conditions, including institutional support. Adomako et al. (2020) discuss how government R&D support can enhance inter-firm cooperation by bridging institutional gaps, indicating that effective collaboration can thrive in favorable policy environments. This highlights that collaboration is often a response to external economic realities.

Geographic proximity also plays a crucial role in the effectiveness of collaborative interactions. Jørgensen et al. (2011) highlight the importance of physical proximity in fostering innovation and cooperation, pointing to the need for companies to

strategically consider their location strategies when engaging in collaborative ventures. In contrast, the concept of boundary-spanning networks, as outlined in Andrade Rojas et al. (2018), suggests that cooperation between various geographic and organizational boundaries can produce innovative results, thus enhancing overall firm performance.

As firms increasingly engage in collaborative ventures, the role of knowledge spillovers becomes pivotal. Montoro Sánchez et al. (2011) investigate how innovation and collaboration are interlinked within science and technology parks, demonstrating that inter-organizational networks can enhance firms' innovation capacity in these ecosystems. It indicates that fostering these collaborative frameworks helps in tapping into shared knowledge, conducive to innovation.

The evaluation and performance measurement of collaborative networks is another critical area of discussion. Goldstein and Butler (2010) note that aligning assessments with clearly defined performance indicators is vital, allowing stakeholders to measure progress and adapt strategies accordingly. Without a systematic approach to tracking output, networks risk losing sight of their objectives, leading to diminished engagement and ineffective resource allocation. Therefore, performance metrics need to be developed collaboratively and routinely assessed to ensure ongoing alignment with organizational goals.

The open innovation paradigm emphasizes the enhancement of an organization's innovation potential through external collaboration Chesbrough (2003), particularly by leveraging external reservoirs of creativity and knowledge Enkel et al. (2009), as well as by facilitating learning through knowledge transfer processes (Secundo et al., 2019). Moreover, the Horizon 2020 initiative underscores the importance of fostering international and interorganizational cooperation (Novitzky et al., 2020; Veugelers et al., 2015). The critical role of international collaboration has also gained recognition by the European Union (EU) and worldwide, particularly in coordinating the response to the COVID-19 pandemic (Cai, 2023).

The field of corporate collaboration network research faces numerous methodological challenges, particularly when studies focus on the European context. These challenges arise from the intricacies of managing data, establishing reliable metrics, and understanding the human and organizational dynamics at play in collaboration networks.

One of the most prominent challenges in studying corporate collaboration networks is the selection and collection of appropriate data. Researchers often rely on SNA methods, which can provide insight into collaboration patterns, but also have significant limitations on the completeness and representation of the data. For example, variations in access to network data can lead to biased conclusions. In corporate settings, where some collaborations may be disclosed, while others remain confidential due to competitive concerns or proprietary strategies, obtaining a complete data set becomes problematic (Joy et al., 2024; Ozer et al., 2013).

In addition, collaboration networks have been investigated at multiple analytical levels. At the country level, sources such as patent databases De Prato and Nepelski (2014), publication and citation databases Guan et al. (2016a) and Guns and Wang (2017), and international trade databases Guan et al. (2016b) and Liu et al. (2020) have been utilized. Regional-level analyzes have been mainly based on publication databases Chuanming et al. (2017) and patent databases De Noni et al. (2018), while at the city level, publication databases have been used as the primary data source (Guns and Rousseau, 2014). At the organizational level, studies have predominantly utilized publication databases Chang and Huang (2013), Han et al. (2014), Huang et al. (2018), Lande et al. (2020), and Lee et al. (2012) and patent databases Chen et

al. (2021). Although most of the research has focused on a single database, some investigations have integrated multiple sources-such as Patstat and Orbis Cséfalvay and Gkotsis (2022), Mahnken and Moehrle (2018), and Tarasconi and Menon (2017) or a combination of Orbis and Cordis.

Another methodological hurdle encountered in corporate collaboration research is the definition and operationalization of concepts such as collaboration and innovation performance. Different scholars utilize varying frameworks to measure these constructs, which can lead to inconsistencies in research outcomes and difficulty in drawing generalizable conclusions. Studies investigating the impact of collaboration on organizational results often find that the characteristics of the teams involved can lead to divergent results depending on how one defines collaboration (Kumar and Operti, 2023). This inconsistency can challenge the synthesis of results in different studies and ultimately obscure the true effects of collaboration on innovation and corporate performance (Binte Azhar et al., 2019; Shi and Xiao, 2024).

In the European context, external variables such as regulatory environments and cultural differences must also be factored into the analyzes. Differences in governance structures, market dynamics, and institutional contexts between various European countries can complicate cross-country comparisons and analyses of collaborative outcomes (Ali et al., 2024; Magnusson and Werner, 2022). Researchers must therefore consider these contextual factors when designing studies, leading to increased methodological complexity as they navigate these diverse landscapes (Demir and Lukeš, 2024). In addition, the influence of various stakeholder groups adds another layer, as collaborations involving Non-Governmental Organization (NGO)s, government entities, and private firms often have different objectives and success measures (Ali et al., 2024; Tian et al., 2021).

Another aspect posing challenges is the evaluation of collaboration models through which firms engage with each other, particularly in the domain of corporate-startup partnerships. Evaluating the effectiveness of such models and understanding the mechanisms through which they affect organizational transformation present significant methodological difficulties. Research in this area often struggles with the identification of suitable metrics to effectively evaluate outcomes (Rigtering and Behrens, 2021; Steiber and Alänge, 2020). Although qualitative methods can provide deep insight into participants' experiences and perceptions, they may lack the generalization needed for greater claims regarding the effectiveness of these models (Steiber, 2020).

Research approaches must also consider the impact of network structure and composition, particularly how attributes such as density, centrality, and heterogeneity influence collaborative effectiveness. High-density networks can provide robust communication channels, but can also lead to redundancy and information overload among partners, complicating performance outcomes. Analyzing such configurations requires sophisticated modeling techniques, which can be resource-intensive and require considerable expertise in network analysis (Jiang et al., 2019; Ozcan and Islam, 2014).

The data related challenges are addressed within this work as creating the most comprehensive database which was not identified in the recent literature. Based on the literature review performed, there has yet to be an integration of corporate data (Orbis), collaboration data (Cordis), patent data (Patstat), or economic background data (Eurostat) at the regional or organizational level. However, without the integration of these varied datasets, the precise prediction of collaborations remains unattainable. More details are mentioned in Sections 3.2 and 3.2.2.

Taking into account the challenges mentioned related to the different measures of success, the type of entities, and also the struggle to identify metrics, H2020 cooperation was investigated in this study in which the measures were already established as the entities already participated successfully in subsidized projects. The type of companies was also diverse and the information was collected and included in the research database used (see Sections 3.1 and 3.2.2).

The network structure-related challenges were also experienced during this work and this is one of the main reasons why the large amount of different methods had been used to properly analyze the collaboration network and not to misinterpret any of the outcomes. All relevant network parameters were calculated and considered together with showing the different methods with benefits and further improvements. The related content is mentioned in Sections 3.3.2 and 3.3.2.

In conclusion, the promise of collaboration networks is evident in various sectors, but the challenges accompanying their formation and management cannot be overlooked. The complexity associated with governance dynamics, the integration of diverse technological systems, and the evaluation of organizational capacity for collaborative performance tasks are constantly evolving. Fostering environments of mutual trust and establishing clear communication channels are essential in transcending obstacles to foster meaningful and effective collaborative endeavors.

2.5 Application of null models in economic network analysis

In this section, the main aim and benefit of using null-models in network research area is summarized and the most referenced and used null-model types are mentioned in the recent literature.

The application of null models in economic networks serves as a robust framework for evaluating the significance and structure of these networks. These mathematical constructs allow researchers to establish baseline expectations against which observed data can be compared, thus facilitating the identification of meaningful patterns and relationships within economic systems.

Null models, particularly configuration models, are widely employed in network analysis by fixing the degree sequence of the nodes while randomizing the connections. Fosdick et al. (2018) noted that random graph null models can elucidate whether the observed properties of economic networks are meaningful or merely the product of their degree sequences. Foster et al. (2010) elaborated on the utility of degree-based null models to assess structural properties in networks, showing how these models can be integral to the estimation of statistical significance in empirical data. Such models have become standard tools for analyzing complex economic networks ranging from trade relationships to market interactions.

Importantly, the choice of null model can significantly influence the interpretation of structures observed in economic networks. Ren et al. (2020) assessed the nesting of world trade networks and compared these structures with random networks generated using null models, revealing that real networks often exhibit arrangements that defy the expectations of randomization in terms of their hierarchical structure and connectivity. This highlights the ability of appropriate null models to uncover anomalies and significant relationships that may not be evident at first glance.

In addition, the methodology for generating null models has expanded to encompass various types of economic interactions. Li et al. (2015) discussed generating

null models for large-scale networks through edge rewiring, which preserves certain statistical properties of original networks while modifying others, thus allowing fine-grained analysis of network topology. Such advancements allow researchers to better understand the complexities of economic interactions by providing context and clarity against which empirical anomalies can be judged.

Another intriguing area of study is the notion of hyper-null models discussed by Zeng et al. (2023), which involve multidimensional aspects of networks. These models are vital in analyzing economic networks characterized by more complex relationships and interactions, allowing analysts to probe deeper into the implications of connectivity beyond simple pairwise interactions. By using refined null models tailored to specific types of economic data, researchers are better equipped to distinguish between random chance and genuine structural features of economic connections.

Without claiming completeness, three of the most referenced null-models with a brief explanation and the main benefits and challenges of them are mentioned below highlighting the wide range of the available models together with the importance of them. As an addition, one of the most beneficial for current work, the Gravity-based null model is discussed separately in Section 2.6.

2.5.1 Erdős-Rényi Random Graph Model

The ER-model represents one of the foundational null models in network science (Erdős and Rényi, 1959). There are two main variants: $G(n, m)$, which randomly selects m edges from all possible connections, and $G(n, p)$, which establishes each possible edge with independent probability p. In economic networks, these models serve as baseline comparisons to detect non-random structures in empirical data. For example, when analyzing interbank lending networks, ER-models help identify whether observed clustering of financial relationships exceeds what would be expected by chance (Boss et al., 2004).

Advantages: The primary strength lies in mathematical tractability, allowing analytical derivations of expected network properties (Jackson, 2008). The model provides clear thresholds for the emergence of giant components and connectivity, that can be interpreted as critical points for system-wide economic integration or the spread of contagion (Albert and Barabási, 2002). ER-models offer straightforward computational implementation and serve as an intuitive benchmark against which more complex economic network formation processes can be evaluated (Schweitzer et al., 2009).

Challenges: The uniform probability assumption contradicts most economic interaction mechanisms, where preferential attachment, strategic behavior, and historical paths dependencies play crucial roles (Barabási and Albert, 1999b). ER-models generate Poisson degree distributions, failing to capture the heavy-tailed distributions typically observed in trade networks, financial relationships and other economic systems (Schweitzer et al., 2009). They also produce networks with negligible clustering coefficients in the large network limit, contrary to the significant triadic closure seen in real economic networks, particularly those influenced by trust relationships or geographic proximity (Jackson, 2008).

2.5.2 Configuration models

Configuration models preserve the degree sequence of an empirical network while randomizing connections, providing a more sophisticated null model than the ER-model (Newman, 2010a). Implementation typically involves cutting all edges to create "stubs" and then randomly rewiring these stubs while maintaining each node's original degree. In economic networks, this approach acknowledges the fundamental heterogeneity in the connectivity of actors. Some firms maintain numerous trading partners, while others engage with only a few, while testing whether other structural characteristics arise from processes beyond the degree distribution (Squartini and Garlaschelli, 2011).

Advantages: By preserving heterogeneous connectivity patterns, configuration models provide a more realistic baseline for economic networks where size and activity distributions are highly skewed (Mastrandrea et al., 2014). They enable researchers to identify patterns beyond what can be explained by degree sequence alone, such as core-periphery structures in production networks or assortativity in international trade (Serrano and Boguñá, 2003). Configuration models can be analytically approached using maximum entropy principles, establishing connections to statistical physics frameworks that allow rigorous hypothesis testing (Squartini and Garlaschelli, 2011). Advanced versions can maintain in-degree and out-degree sequences, crucial for directed economic networks such as supply chains or investment flows (Newman, 2010a).

Challenges: Standard implementations may generate multi-edges and self-loops, requiring additional constraints or rejection sampling techniques that can become computationally prohibitive for large economic networks (Squartini and Garlaschelli, 2011). While preserving degree sequences, these models typically do not capture transitivity, reciprocity, and community structures that often arise from institutional similarities or geographic clustering in economic systems (Squartini et al., 2015). The models assume independence between edge formations conditional on degree constraints, overlooking strategic complementarities or substitution effects common in economic decision-making (Mastrandrea et al., 2014). Implementations using sampling approaches may not fully explore the configuration space, potentially biasing statistical inferences about the formation processes of economic networks (Artzy-Randrup et al., 2005).

2.5.3 Exponential Random Graph Models

Exponential Random Graph Model (ERGM)s define probability distributions over networks based on configurations such as edges, stars, triangles, and other local structures (Robins et al., 2007). These models can be expressed as follows:

$P(G) = \frac{1}{Z} \exp\left(\sum_k \theta_k s_k(G)\right)$ where $s_k(G)$ represents the network statistics, θ_k are parameters, and Z is a normalizing constant. In economic applications, ERGMs can incorporate both endogenous network dependencies (e.g. reciprocity in trading relationships) and exogenous covariates (e.g. geographical distance, institutional similarity) (Lusher et al., 2013). They have been applied to interbank lending networks, international trade, corporate ownership structures, and supply chains (Amini et al., 2013).

Advantages: ERGMs offer unparalleled flexibility in simultaneously modeling multiple network formation mechanisms, allowing researchers to test competing

economic theories (Robins et al., 2007). The models provide a principled statistical framework grounded in maximum entropy principles, connecting to established econometric approaches (Lusher et al., 2013). ERGMs can incorporate node attributes such as firm size, country Gross Domestic Product (GDP), edge attributes like trade volume, loan terms, and structural dependencies (transitivity, preferential attachment), making them comprehensive tools for economic network analysis (Snijders et al., 2006). The explicit specification of parameters provides interpretable coefficients that quantify the importance of different economic mechanisms in network formation (Hunter et al., 2008). Advanced specifications can model dynamics through temporal dependencies, capturing how economic networks evolve in response to changing conditions (Hanneke et al., 2010).

Challenges: ERGMs suffer from well-documented degeneracy issues where the model produces near-empty or near-complete networks, particularly problematic when modeling sparse economic networks with complex dependencies (Handcock, 2003). Estimation becomes computationally prohibitive for large networks, limiting applications to many real-world economic systems that involve thousands or millions of actors (Snijders et al., 2006). Markov Chain Monte Carlo methods used for estimation may converge slowly or fail to converge, especially when models include higher-order dependencies (Lusher et al., 2013). Interpretation of parameters becomes challenging when multiple interdependent effects are included, making it difficult to isolate the impact of specific economic mechanisms (Hunter et al., 2008). The models typically assume network equilibrium, which may not hold in rapidly evolving economic systems, particularly during crises or significant policy changes (Snijders, 2011).

In general, null models in economic networks are indispensable for discerning the intricacies underlying networked economic interactions. Their application aids in the quantitative assessment of network properties, enhances the understanding of complex relationships, and ultimately provides a clearer rationale for the observed phenomena in economic systems.

Based on the benefits and challenges mentioned for the different models, in the dissertation, the ER-model was used as the standard baseline model for the investigation of the ownership network. The other model was the configuration model, but the main beneficial one was the gravity-based one, which is discussed separately in the next, 2.6 Section because of the higher importance of it within this work.

2.6 Gravity model

The gravity model serves as a foundational framework for analyzing economic networks, particularly as a null model, which provides a baseline against which actual economic interactions can be compared. Originally inspired by Newtonian physics, the gravity model posits that the interaction between two economic entities (like countries or regions) is directly proportional to their economic sizes (usually measured by GDP or population) and inversely proportional to the distance between them. This model has been widely applied and modified in various studies to suit specific economic contexts, particularly in assessing trade, tourism, and urban mobility networks.

The gravity model is commonly used for understanding the structural characteristics of economic networks due to its inherent adaptability and robustness. For example, the model effectively captures urban mobility patterns and regional economic interactions. Tang et al. (2022) discuss the applicability of the gravity model in

estimating urban mobility networks in geographic contexts such as the Guangdong-Hong Kong-Macao Greater Bay Area, thus supporting its relevance in economic pattern analysis (Tang et al., 2022). Furthermore, research by Gan et al. (2024) indicates that the modified gravity model can assess economic links in tourism networks, providing significant information on city interactions in border economies.

Furthermore, the gravity model's application extends beyond mere trade analysis; it also encompasses multilayer analyses in finance and macroeconomic dynamics. Sharma et al. (2019) illustrate how gravity equations can delineate relationships within multilayered economic networks, helping to elucidate complex interdependencies in financial systems. This multifaceted applicability demonstrates the significance of the gravity model as a framework for evaluating economic networks. Studies by Xie et al. (2021) have also shown how the gravity model can be adapted to incorporate temporal variations, effectively predicting tourism flows.

Using the gravity model as a null model allows researchers to establish what constitutes "normal" economic interaction patterns. When actual data deviates significantly from these predictions, it reveals anomalies. For example, Reyes et al. (2014) highlights the model's capability to simulate baseline trade patterns under various regional trade agreements, helping to analyze trade effects through network approaches. By comparing actual trade flows with predictions made by gravity models, researchers can identify significant deviations, shedding light on the impacts of external variables such as policy changes or economic shocks (Fagiolo, 2010).

However, it is crucial to acknowledge the limitations inherent in the gravity model. The model simplifications can obscure complex dynamics in economic networks, as highlighted by Chávez-Bustamante et al. (2023), who emphasize the importance of considering noneconomic factors and indirect effects in trade and migration flows. Hence, while the gravity model provides a valuable starting point, subsequent analyses often require more nuanced frameworks that account for the complexities of real-world interactions.

In summary, the main challenges and benefits can be summarized as follows.

Advantages: Gravity models incorporate economic theory directly into network null models, providing theoretically grounded expectations about connection patterns (Squartini and Garlaschelli, 2018). They excel in capturing spatial dependencies in economic networks, acknowledging that distance (whether geographical, cultural, or institutional) fundamentally shapes economic interactions (Fagiolo et al., 2010). The models can be extended to include multiple dimensions of distance and various barriers to interaction, such as tariffs, language differences, or colonial ties (Head and Mayer, 2014). They establish direct connections with established econometric literature, allowing researchers to use advanced estimation techniques and diagnostic tools (Santos Silva and Tenreyro, 2006). Gravity models provide explicit counterfactuals for policy analysis, such as estimating the effects of trade creation and diversion of economic integration agreements (Baier and Bergstrand, 2007).

Challenges: Gravity models require detailed node attribute data, which may not be consistently available, especially for subnational entities or developing economies (Dueñas et al., 2017). The standard log-linearized estimation approach struggles with zero flows, which are prevalent in sparse economic networks (Santos Silva and Tenreyro, 2006). These models often fail to capture complex nonlinear relationships and interaction effects between different economic factors (Anderson, 2011). They typically assume independence between observations after controlling for size and distance, overlooking network dependencies such as preferential attachment or strategic complementarities (Squartini and Garlaschelli, 2018). Gravity models

may oversimplify the multifaceted nature of economic relationships that are not determined primarily by size and distance, such as strategic partnerships or political alliances (Head and Mayer, 2014).

In conclusion, the gravity model acts as a critical null model in economic network analysis, allowing researchers to gauge the extent of economic interactions based on fundamental principles relating economic size and distance. Its adaptability and integration into various economic contexts reinforce its utility for modeling not only trade phenomena but also broader economic links. Despite its simplifications, the gravity model remains instrumental for theoretical explorations and empirical applications alike. As discussed in Section 3.3.1, all the challenges are addressed, and therefore all the benefits can be applied. With some minor modification and additional preparation, this model can be suitable for use as the proper one for the investigation of European corporate ownership networks.

2.7 Link prediction in networks

Link prediction in networks involves predicting the likelihood of missing or future links based on the existing structure of the network (Chen et al., 2021). Various methods have been developed to address link prediction challenges in multiple domains - beyond social networks - applications of link prediction extend to fields such as bioinformatics, information retrieval, and e-commerce (Hasan and Zaki, 2011). In general, these methods fall into categories such as similarity-based, probability-based, machine learning-based, embedding-based algorithms (Chen et al., 2021).

Link prediction has been recognized as a critical component in the analysis of scientific collaboration networks, with numerous studies focused on forecasting collaborations between partners, such as authors or organizations. A collaboration recommendation model Chuanming et al. (2017) was developed using information from network neighbors and paths within scientific collaboration networks. An empirical assessment was conducted to assess the effectiveness of the model at the individual, institutional and regional levels. In 2019, eight weighted algorithms were constructed by Wang et al. (2019b) to predict possible scientific collaborations, revealing that algorithms incorporating fusion indicators produced better predictive performance. In their work, two new indicators were proposed and combined with four commonly used metrics: Institutional Document Frequency (IDF) and Institutional Cumulative Cooperation Ratio (ICCR), which served to quantify the similarity of the authors' references when selecting the collaborating institutions. These analyzes were performed at the organizational level.

Subsequently, a method for predicting collaborations between scientists was introduced by Lande et al. (2020), employing a heterogeneous information network model composed of authors and keywords extracted from the Web of Science database. Eight different link prediction approaches were evaluated by Chen et al. (2021) to investigate partner selection within the China interorganizational patent cooperation network. The challenge of selecting appropriate collaboration partners for innovation was addressed by Qi et al. (2022), who established a framework that integrates topic analysis with link prediction techniques. Their study identified three subcategories of literature-based methods: (1) network analysis/link prediction approaches utilizing single or multiple network indicators, or combining these with external feature indicators derived from bibliographic data; (2) index-based methods relying on bibliographic information; and (3) approaches employing Natural

Language Processing (NLP) techniques to analyze the intrinsic content of the literature. In their research, a link prediction method that uses a fusion network of authors and patentee organizations was applied to predict potential collaborations.

Similarity-based techniques leverage local structures around the nodes when predicting potential links (Liu et al., 2016; Sun et al., 2017; Xu and Yin, 2017). Well-known examples include the Common Neighbor Method (CN) method, Jaccard coefficient, and Adamic-Adar index. The fundamental premise of similarity-based algorithms is that nodes with a greater number of common neighbors are more likely to be connected. The CN method is intuitive, where the prediction score is directly proportional to the number of mutual connections between two nodes, which is computationally less intensive (Ahmad et al., 2020; Zhang et al., 2016b). The Jaccard coefficient, on the other hand, calculates similarity as the size of the intersection divided by the size of the union of the neighborhood sets of two nodes, which ensures normalization by the total degree of connections involved (Singh, 2023; Zhang et al., 2016b). Although straightforward and easy to compute, similarity-based algorithms have limitations, particularly when it comes to large graphs, where they may neglect global structural features that could be significant.

A notable downside to similarity-based approaches is their reliance on local structure, which may overlook the importance of broader connectivity patterns. For example, while the CN method thrives in dense networks, its efficiency is reduced when applied to sparse networks where connections might be less evident (Bi et al., 2024). Furthermore, these methods may produce biased results due to structural regularities in real-world networks, possibly leading to overprediction in high-degree nodes and underprediction in low-degree nodes (Dimitriou and Karyotis, 2024; He et al., 2024).

Probability-based methods estimate the existence of potential links by constructing statistical models that quantify the probability of link formation (Chi et al., 2019; Yang et al., 2014).

In recent years, machine learning methods Li et al. (2018), Mohan et al. (2017), and Zhang et al. (2016a) have gained prominence in link prediction tasks due to their ability to model complex relationships and patterns that simple similarity metrics might miss. These include supervised methods such as logistic regression and various ensemble techniques, as well as unsupervised methods such as clustering approaches. Machine learning algorithms can incorporate a larger set of features, such as node attributes and hyperparameters that define local link structures Sulaimany et al. (2017) and Wang et al. (2020). For example, deep learning approaches utilizing neural networks can automatically extract features from the network structure and provide high prediction accuracy in various contexts (Deng et al., 2023; Parisi et al., 2018). Metaheuristic approaches Bastami et al. (2019), employ intelligent computational frameworks for link prediction. Due to their ability to predict links between unconnected nodes, these methods are applicable to partner selection processes in patent collaboration networks (Chen et al., 2021).

However, the application of machine learning to link prediction also has its drawbacks. One of the significant challenges lies in the requirement for large labeled datasets to train such models effectively. In situations where labeled data are sparse or unavailable, learning models can perform poorly, leading to a decrease in predictive accuracy. Moreover, the complexity of tuning machine learning algorithms may cause them to be less interpretable compared to simpler and more traditional methods (Moutinho et al., 2024).

Embedding techniques offer another powerful framework for link prediction by

representing nodes in a continuous vector space where geometric proximity corresponds to relational proximity. These methods, such as node2vec and DeepWalk, use random walks or neighborhood sampling to learn low-dimensional representations of nodes that capture their topological characteristics within the network. This approach has proven advantageous, especially in large networks, as it can scale to millions of nodes while still preserving the properties of the linkage (Xin and Zhao, 2009; Yuliansyah et al., 2020).

Embedding methods also exhibit the ability to generalize well to new or unseen nodes, a critical aspect in networks where information might evolve over time. However, challenges include the potential for overfitting, particularly in networks with sparse connections, as well as difficulties in capturing dynamic changes in network structure if the embeddings are not updated regularly (Gu et al., 2023; Wang et al., 2019a).

Complex real-world networks often exhibit nuances that require hybrid models that combine various techniques for more effective link prediction. For example, methods that synergize both local similarity indicators and global structural properties have been proposed to balance the limitations of purely local approaches. By integrating insights from similarity-based approaches and machine learning, these hybrid models can produce more robust predictions (Dimitriou and Karyotis, 2023).

Furthermore, advances in GNN introduce new para-digms where the message passing framework allows for community-based link prediction, leveraging both node features and network topology to improve prediction accuracy (Huang et al., 2016; Liu et al., 2019). GNNs have been widely adopted for link prediction in complex networks due to their ability to effectively capture the underlying structural information of the network. Node embeddings can be learned and links predicted based on both the graph topology and the patterns inferred from existing connections. Through this approach, accurate predictions about potential links between nodes in the network can be made, even in scenarios where only incomplete information is available (Zhou et al., 2020).

KGE techniques enable the embedding of entities and relations from knowledge graphs into low-dimensional vector spaces, thus facilitating tasks such as link prediction and knowledge discovery (Ge et al., 2024). Path-based algorithms utilize semantic paths between entities within knowledge graphs to predict missing links and infer novel relationships (Rossi et al., 2021).

GCN generalize convolutional neural networks to graph-structured data, allowing the modeling of node relationships in link prediction tasks in various domains, including social networks, citation networks, biological networks, recommendation systems and security applications (Zhang et al., 2019). Promising results in link prediction have been reported for GCNs, attributed to their ability to capture both local and global structural properties of graphs (Zhou et al., 2020). Furthermore, variants such as graph attention networks and graph recurrent networks have emerged as advanced forms of GNNs, demonstrating exceptional performance in a range of deep learning tasks, including link prediction.

However, while hybrid techniques often achieve better performance, their complexity can make them computationally intensive and harder to implement.

Another area of advancement within link prediction is its application to dynamic networks, where the structure evolves over time. Algorithms designed for this purpose account for temporal data, which is critical in social networks or biological communication systems, where relationships change frequently. Temporal modeling helps identify not only recent trends but also predict future connections based on observed changes in the network (Bayrak and Polat, 2018; Li et al., 2021).

It should be noted that while these dynamic approaches facilitate timely accuracy, they may struggle with handling rapidly changing data or inferences when historical data are less representative of the current topology (Dong et al., 2013; Zhang et al., 2023). This shows a crucial trade-off between accuracy and adaptability in models used for dynamic link prediction.

In conclusion, link prediction remains a vital area of research with various methodologies that include similarity-based, probability-based, machine learning, and embedding strategies. Each approach presents its own advantages and limitations, with ongoing research aimed at improving predictive accuracy and computational efficiency. The choice of method greatly depends on the specific characteristics of the network in question, including its size, density, dynamism, and the nature of the relationships involved.

2.8 Multilayer networks

Multilayer networks represent a significant advance in the study of complex systems, enabling researchers to visualize and analyze intricate relationships in diverse domains. Using multiple layers of interconnected networks, these networks facilitate a deeper understanding of systems in which interactions among nodes are multidimensional rather than one-dimensional. This framework has been applied across various fields, underscoring the versatility and adaptability of multilayer networks in contemporary research.

The historical development of multilayer networks stems from the recognition that many systems comprise various types of interactions occurring simultaneously among interconnected entities. Early studies in network science focused on single-layer networks, which assumed uniformity in interactions among elements. As the limitations of these simplified models became apparent, the multilayer approach emerged, emphasizing that interactions can manifest across different layers and, thus, enrich the understanding of system dynamics (Hammoud and Krämer, 2020; Pilosof et al., 2017; Ye et al., 2021).

A notable contribution to the evolution of multilayer networks is provided by Wang et al. (2015b), who articulated the interdependencies inherent in multilayer systems through evolutionary game theory. This perspective demonstrates how multilayer networks can model complex social phenomena, such as cooperation and competition among interconnected groups, enabling a deeper understanding of social dilemmas. These findings highlight the importance of multilayer frameworks in elucidating interaction mechanisms that would remain obscured within a unidimensional framework.

The concept of multilayer networks arises from the need to represent systems where single-layer networks do not capture complex interactions among different entities or processes. A multilayer network consists of multiple interconnected networks, each denoting a different type of relationship or interaction among the same set of nodes such as individuals, organizations, or systems (Domenico et al., 2013; Wider et al., 2016). Unlike single-layer networks that often simplify real-world dynamics, multilayer structures provide a framework for understanding cross-domain interactions and dependencies, which are crucial in decision-making processes within economics (Chen et al., 2017a; Jiang and Liang, 2024; Li et al., 2023).

Multilayer networks can effectively model cooperation between various economic agents. For example, collaborations in trade relationships can be examined through distinct layers representing different types of exchanges (monetary, goods, services)

between participants (Chen et al., 2017b; Wang et al., 2022). By analyzing these interactions at multiple levels, researchers can identify synergies that might be overlooked in traditional models.

Moreover, methodologies for time series forecasting using multilayer perceptrons have gained prominence. Xin and Zhao (2009) explored a Monte Carlo-based algorithm to construct multilayer neural networks dedicated to time series forecasting. Their results indicated improved predictive performance through sophisticated network architectures.

Furthermore, multilayer networks provide a robust framework for analyzing temporal dynamics. Iwayama et al. (2012) and Lacasa et al. (2015) demonstrated the transformation of multivariate time series data into multilayer representations, enabling the extraction and analysis of complex patterns across time and dimensions. This capability is pivotal in fields ranging from physics to social sciences and finance, highlighting the versatility of multilayer networks in capturing system behavior.

The inherent complexity of multilayer networks may pose computational challenges, especially as the number of layers increases. Analyzing such intricate interdependencies often requires advanced algorithms and greater computational resources, which may not be readily available Chen and Zhu (2016) and Shao et al. (2015).

Collecting and integrating data from multiple layers can be difficult. Economic data often reside in disparate sources, making it difficult to ensure consistency and completeness (Chen et al., 2017b; Zhang et al., 2022). Without robust data integration methodologies, the potency of multilayer networks may be undermined.

Economic conditions are not static. They evolve based on reforms, policy changes, and external shocks (for example, pandemics). Capturing such dynamism within a multilayer network is challenging, as it requires continuous adaptation and reevaluation of structural relationships between layers (Sun et al., 2024; Zanin, 2015).

The identified challenges are also considered for the application of multilayer network representation in this study. The increase in complexity with an increase in the number of layers was not a problem, as in this case the different layers represented the different years in the investigated time frame and there were no connections between the layers (see Section 3.3.1). The data relevant and the economic parameter changes are handled properly with the creation of the research database (see Sections 3.2 and 3.2.1).

The development and application of multilayer networks have had a transformative impact across domains such as biomedicine, ecology, cognitive science, industrial engineering, and so on. This network structure integrates multiple layers of interaction, allowing a comprehensive analysis of the complexities inherent in real-world systems. As the field evolves, multilayer networks are poised to play an increasingly prominent role in advancing research and practical applications aimed at understanding and managing complex systems.

2.9 Horizon 2020

This section shows a brief introduction of H2020 reflecting the main goals and challenges of it. This summary helps to understand the importance of the program and the magnitude of it's importance in the European research collaboration.

H2020 represents the most ambitious and expansive research and innovation framework initiative undertaken by the European Union (EU). Launched on January 1, 2014, and finalized at the end of 2020, it operated with a considerable budget of nearly € 80 billion, which was the largest allocation in the history of EU funding programs dedicated to research and development (Annette, 2025; Hogan, 2017; Sharp, 2019). As the eighth framework program to succeed its predecessor, it built on the strategic objectives established throughout previous iterations while simultaneously adapting to the evolving global context of research and innovation. The overarching purpose of H2020 was to address critical societal challenges, enhance the global competitiveness of the EU, and ensure that strategic sectors are nurtured within the vibrant research ecosystem of Europe (Puślecki, 2016; Sharp, 2019).

The program was structured around three principal pillars: Excellent Science, Industrial Leadership, and Societal Challenges. The first pillar, Excellent Science, focused on reinforcing the EU's scientific leadership through funding for basic research, promoting curiosity-driven scientific inquiry, and attracting global talent to research hubs across Europe. Significant initiatives included the European Research Council (ERC) grants that provided substantial support for cutting-edge research projects. The second pillar, Industrial Leadership, aimed to stimulate the growth of world-class innovation by nurturing Small and Medium sized Enterprise (SME)s and facilitating public-private partnerships. By connecting researchers with industry leaders, this pillar aims to advance the commercialization of research output, thus enhancing economic growth and job creation (Saletti et al., 2020).

In addition, the third pillar, Societal Challenges, dealt with pressing global issues such as climate change, health, energy, and security, thus directly contributing to societal welfare. The programs in this pillar sought to mobilize research and innovation to find solutions to these challenges, thus fostering a direct connection between science and public policy (Georgiadou, 2018; Saletti et al., 2020; Uhrig, 2019). H2020 also placed significant emphasis on promoting interdisciplinary collaboration and participation with stakeholders, ensuring that the results of funded research translated into meaningful social impact (Sharp, 2019; Uhrig, 2019).

H2020 was not merely a funding framework; it sought to create an environment conducive to innovative approaches aimed at social transformations. The program included significant investment in addressing the challenges of climate change and resource efficiency as outlined in the Paris Agreement, focusing particularly on renewable energy systems (Georgiadou, 2018). This commitment was reflected through various funded projects that aimed to develop sustainable technologies operating within Europe's future green economy (Georgiadou, 2018; Hogan, 2017). The role of enhanced partnerships, collaboration, and the integration of research within policy was consistently highlighted as key elements in achieving the goals set out in the framework (Saletti et al., 2020).

The integration of gender equality and diversity also appeared prominently in H2020, reflecting the larger commitment of the EU to ensure inclusion and tackle disparities within the research landscape. The implementation of the program included specific guidelines and initiatives to promote gender inclusion in all funded projects, although challenges remained with respect to achieving equitable participation (Vida, 2020). By weaving in considerations of innovation, collaboration, and inclusivity, Horizon 2020 aimed to ensure that research not only advanced scientific knowledge but also supported social transformations that mirror the diverse needs of the European populace.

In evaluating its effectiveness, an interim report in 2017 reflected on the impact of H2020 through quantitative analyzes and qualitative feedback from stakeholders

(Annette, 2025; McCarthy, 2018). The evaluation suggested that while many initiatives effectively promoted innovation and collaboration, more improvements were needed to improve citizen engagement and raise public awareness about the results of EU-funded research (McCarthy, 2018). The emphasis on communication and accessibility of results was highlighted as an area that is ready for development, particularly to ensure that scientific advances benefit all layers of society (Uhrig, 2019).

Furthermore, the extensive evaluation process highlighted the need for continuous adaptation of funding mechanisms to align with the changing landscape of science and technology and to respond more effectively to emerging societal challenges. The challenges posed by rapid technological advances and changes in global priorities required future frameworks to build on the lessons learned during H2020 (Barbu and Niţă, 2025; Saletti et al., 2020).

As H2020 transitioned to its successor program, Horizon Europe, the successes and challenges encountered will inform future policy decisions and funding strategies. Horizon Europe aims to strengthen the EU's position as a leader in global research and to foster a more resilient European economy. By addressing previous criticisms and successes of H2020, the new framework looks set to fine-tune its approach to maximize the potential for research and innovation in all member states (Annette, 2021).

In general, Horizon 2020 has been instrumental in shaping the European research agenda, promoting scientific excellence, and empowering innovative solutions that address current and future social challenges. The impact of the framework can be seen in its facilitation of effective public-private partnerships, investment in groundbreaking research, and commitment to inclusion within its processes (Annette, 2021; Fałdowski and Nepelski, 2018; Sharp, 2019). As the EU moves forward, the lessons gleaned from H2020 will undoubtedly serve as a foundation upon which to build a more innovative and cohesive European research landscape that continues to inspire global collaboration and progress.

2.10 Research assumptions

By revisiting the research questions established in Section 1.5 and conducting a critical review of the findings and connections with the already available and above mentioned literature, it becomes possible to formulate the corresponding research assumptions. The three research assumptions are as follows.

> **RA1:** Gravity-driven economic principles dominate ownership network formation, with gravity-based economic null model predictions reflecting real-world investment flows more accurately than topology-only models.

> **RA2:** Administrative borders create structural breaks in ownership networks independent of geographic proximity, persisting across temporal layers.

> **RA3:** Machine learning techniques including generic and non-generic approaches can be beneficial for improving the prediction of the connections in the collaboration network of the Horizon 2020 Programme and with the proper model, the influential factors can also be identified.

Chapter 3

Data and Methods

In the following section, the data and methods that were used in this work are introduced. The exact sources of the data with details about the structure and content inside are mentioned together with the used databases, and the main purpose they are used for is also mentioned. In addition, the main data cleaning and processing steps that were performed to make the data usable for the investigation are explained. In the second part, the methods used to investigate networks based on the data used are introduced, mentioning the reason why the mentioned method was used together with the challenges and benefits of all the methods mentioned in the section.

3.1 Data Sources

As already mentioned, for an investigation like this, a large amount of data is needed, which is available nowadays only from different and in most cases heterogeneous sources and structures. In this Section the used data sources and employed data are introduced together with the purpose behind them.

In Figure 3.1, the main data sources can be seen, which were included in a comprehensive research database.

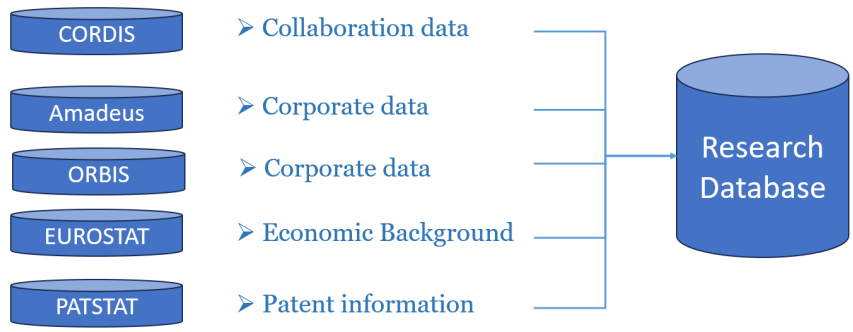


FIGURE 3.1: Input sources for the research database

3.1.1 Cordis database

The Cordis[1] database is provided by the Community Research and Development Information Service, which contains information about the Framework Programmes

[1] https://cordis.europa.eu/

in the EU. Although the main focus was on the Horizon 2020[2] program in the EU, the participation information of the FP7[3] program participation information was also integrated. This database contains data about projects and contributors to projects. For this work, a limited set of data was available from the complete Cordis dataset as part of the data property of the research group involved in this dissertation.

The dataset used for the H2020 participation information contains more than 25 thousand earned projects and more than 150 thousand participants who were all part of the H2020 Framework Progamme. The ratio of European entities was 86% as within the participants in projects, there were not only European entities, but also entities from other countries outside the EU. This dataset contains all the important data about the projects such as start and end dates of them, description of the project goals, funding scheme, earned contributions by projects, full budget of them, etc., and moreover also the organizations main information is available inside. The main information about the organizations was their names, based on what it was possible to connect with other datasets.

3.1.2 Amadeus database

The Amadeus database, just like Orbis[4], is a database from Bureau van Dijk which was acquired by Moody's[5]. Amadeus was retired on 30 November 2022 but before its closure, a comprehensive dataset was gathered by the research team for use in this work. The data were available in raw format in thousands of excel files separately, which had to be processed to make them usable for further work. The data preparation task was mainly done using specific C# software codes which were written manually as the data sources in excel files were quite specific.

This data set was the basis for investigating the investment network that contained the ownership information about European companies with the mother-and-daughter relationship information on the company level.

3.1.3 Orbis database

As mentioned earlier, this dataset is available as the successor of Amadeus by Bureau van Dijk company which was acquired by Moody's company and it is a commercial one for which the University has a subscription and all the data needed for this work was able to be reached and downloaded. The database is the most comprehensive dataset which is available in Europe and contains information about more than 100 million organizations. The database stores data on organizational level and has a huge amount of different variables for companies and institutes all around the globe. As is well known also in the literature, the large and very large companies are overrepresented in the database but it contains also a big amount of smaller entities with an acceptable amount of filled and available information.

The dataset contains all relevant identification data like BvD ID number which is the unique identifier for all entities inside the database. In addition to that there are available also other identification information like Legal Entity Identifier, name of the entity, address with all the details, but also the ownership information.... In

[2]https://research-and-innovation.ec.europa.eu/funding/funding-opportunities/funding-programmes-and-open-calls/horizon-2020_en

[3]https://transport.ec.europa.eu/transport-themes/research-and-innovation/seventh-framework-programme-building-europe-knowledge_en

[4]https://www.moodys.com/web/en/us/capabilities/company-reference-data/orbis.html

[5]https://ir.moodys.com/press-releases/news-details/2017/Moodys-to-Acquire-Bureau-van-Dijk/default.aspx

addition this database also contains for most of the entities the NUTS[6] information, which was also helpful to make territorial based investigation as in some part of this work aggregated information for NUTS 3 level region which is considered in EU as county level area was used.

The most important information for this work was the economical and financial information from this dataset. The main focus was on one hand the subsidiary information, namely which company owns which other ones, on the other hand the key financial information for all the companies and other type of entities investigated to be able to define variables for the prediction tasks as additional independent variables. The detailed information about which of the exact data were used in this work, refer for section 3.2.

3.1.4 Patstat database

The Patstat database[7] is a commercial data source from the European Patent Office (EPO) Worldwide Patent Statistical Database, henceforth called "Patstat", which is considered the most prominent database for patent information around the world. It offers patent data and information in bibliographic structure for more than 100 patent offices in which data can also be found from the nineteenth century. The database is structured into separated tables, and the tables are connected to each other with unique identifiers which are relevant only inside the Patstat dataset. The database itself contains more than 100 million records in total, whose number is continuously growing as time lapses.

The dataset contains filed and accepted patents, trademarks, and industrial designs with a large amount of additional administrative information such as the address of the inventor and exploiter. The 2019 Autumn edition, which contained data for patents until 31st of July 2019, was obtained by the research group. Patstat is updated twice every year and released as Spring and Autumn editions which contains data until end of January and end of July, respectively.

Since Patstat includes only those applications that have been published, and considering that the standard publishing delay typically exceeds 18 months, the number of recorded patents is significantly lower for the years 2017 and 2018.

3.1.5 Eurostat database

The last data source that was used during this work is the Eurostat[8] dataset which contains several different areas of the EU economy and is freely available to anyone. Eurostat ("European Statistical Office"; also DG ESTAT) is a Directorate-General of the European Commission located in the Kirchberg Quarter of Luxembourg City, Luxembourg.

The database contains 33 participating countries which are the 27 EU countries, 3 European Free Trade Association (EFTA) counties such as Iceland, Norway, and Switzerland, and 3 EU candidate countries such as Bosnia and Herzegovina, Serbia, and Türkiye.

Eurostat contains not only economical but also cultural, population-related, industrial, technological, and several other information too. For this study, the most

[6] https://ec.europa.eu/eurostat/web/nuts
[7] https://www.epo.org/en/searching-for-patents/business/patstat
[8] https://ec.europa.eu/eurostat/web/main/data/database

important ones were the territorial GDP [9], the Corruption index (CI),[10] and the Persons at risk of Poverty or social exclusion (Pov) [11] information to extend the explanatory variables for each territory.

The GDP information is available in different datasets from Eurostat. The one mainly used contains only the plain GDP data on NUTS 3 level.

The poverty information is part of the domain 'Income and Living Conditions' from which the one used in the database is the 'persons at risk of poverty or social exclusion'. This information is available from Eurostat on NUTS 2 level. The CI is a composite index based on a combination of surveys and assessments of corruption from 13 different sources and scores and ranks countries based on how corrupt the public sector in a country is perceived. The score of 0 represents a very high level of corruption, and a score of 100 representing a very clean country. This data is published by Transparency International[12] on NUTS 1 level.

3.2 Data Employed

In the following section, the exact data used from the sources already mentioned above are described and an insight is given into how the data preparation, cleaning and pre-processing activities were performed to be able to have a comprehensive research dataset as an input for network related investigations. It is also explained how and why two different datasets were used for the investigations as the collaboration network investigation required not only a different method set, but also a different dataset than the ownership network research.

One of the most time consuming activity within the research preparation was the creation of this comprehensive research dataset. The data curation, database creation, data leaning and pre-processing activities were one of my first challenge during the PhD work.

As mentioned earlier, all the used data sources have different structures, use different internal unique identifications and moreover they are not available in a usual database structured system supported by any database engine but most of them were available in different Microsoft Excel, plain text or Comma Separated Values file (csv) files. One of the first tasks was to consolidate and merge these different datasets into one comprehensive database which supports the further activities. The challenge was not only to merge the different sources but data cleaning also had to be done.

First of all, the proper unique identification value definition had to be done to allow all the databases to be connected. The BvD ID number was chosen, which comes from the Orbis database (Orbis is mentioned from now on, but as Amadeus has the same source, structure, and content basis, all the mentions are relevant also for Amadeus database). The meaning of this column is the Bureau van Dijk identification number, which identifies the different companies, and other types of entities in the Orbis database in an unambiguous way. The reason why this was chosen is the phenomenon identified during the data investigation that there are no other

[9] https://ec.europa.eu/eurostat/databrowser/view/nama_10_pc/default/table?lang=en&category=na10.nama10.nama_10_ma

[10] https://ec.europa.eu/eurostat/databrowser/view/sdg_16_50/default/table

[11] https://ec.europa.eu/eurostat/databrowser/view/ilc_peps11n/default/table?lang=en&category=livcon.ilc.ilc_pe.ilc_peps

[12] https://www.transparency.org/en

unique identifications available in any databases. One idea was that the international tax identifier could be used for this goal but it was found that there are examples in the database where different company names belong to the exact same tax identifier number and for all of these different names there are available different and unique BvD ID numbers but the same tax number. One good example for this phenomenon is the city of Paris related companies which are using the same tax identifier (FR72217500016) with several different names and 2934 different BvD ID numbers.

The Cordis dataset does not contain the BvD ID number and to define these identification values, support was received from Burea van Dijk who made available the missing IDs for most of the entities from the Cordis database. Based on the outcome of this step, the connection between the Cordis dataset's participation table was able to be connected to the Orbis dataset via BvD ID numbers and the project related data could be connected too with the project ID numbers.

Based on the address information in the Cordis database, the NUTS regions were possible to identify for all the relevant entities from Orbis (Orbis also contains NUTS information in the database). The connection between the merged Orbis - Cordis dataset and the Eurostat data was done via the NUTS information as in Eurostat that was the smallest unit for which the data was available. The aim was to have as detailed information as possible in the database and in case some aggregation is needed later, based on the detailed availability, it can be done easily.

The information about the patents from Patstat dataset was also integrated into the database but not in the raw format as it is available directly from the source but on NUTS 3 aggregated level. The aggregated information was stored in the database and the identification of the record was done via the NUTS 3 region identifier.

For all the data cleaning, processing, structuring tasks, several tools like C# software codes (C# is part of Microsoft's .NET framework), Python language scripts with the Pandas library were used to identify missing data, rename and restructure the data from the different sources. All the codes in C# and python were written manually and the results of all steps were double checked for correctness.

After all the data sources were integrated into a database system (mySQL was used), all the data which could be used for the investigation work had to be identified. Two main datasets were created from the database and exported into csv files for use in R, as the methods were implemented in R so RStudio was used as the main modeling tool. In total, one edge and one node file for the ownership investigation were created and in parallel another edge and node file for the collaboration network research.

The simplified schema of the merged research database created for this work is shown in Figure 3.2. There are several one-to-many connections drawn on the Figure as the "one" and of the connection represented with the arrow shape which refers that in the table which is targeted by the edge contains only one item which can contain the referred data, while on the other end of the database connection there is used the usual crow's foot notation from ER diagram standard which stand for the "many" end of the connection and means within this table the same data can be multiplied.

For the investigation of the ownership network, the data from Cordis database was not used as in this investigation the whole European community was in the scope of the study, so Cordis data was only used for collaboration related work.

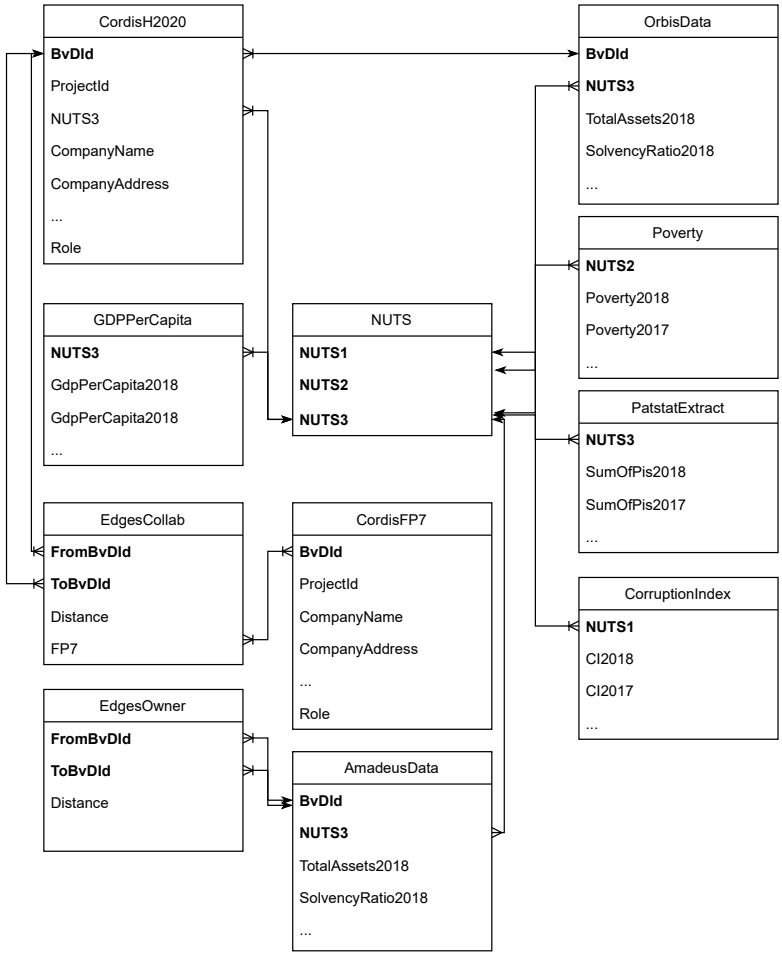


FIGURE 3.2: Simplified Research Database Schema

3.2.1 Used data for ownership network

The investigated time-frame was set between 2010 and 2018 to ensure exclusion of Covid-19 impacts, as the pandemic emerged in 2019 and the company ownership network was analyzed without such influences. The Amadeus database served as the primary data source, containing 23,381,325 companies. Within this dataset, 1,872,272 companies were identified as parent companies or subsidiaries during the examined period. After data cleaning, 1,620,340 companies with verified ownership relationships remained.

These companies were associated with 1,435 NUTS 3[13] regions, which formed the nodes of the temporal network.

87,708 ownership relations were identified between these companies within the studied timeframe, represented as network edges. Connections with both ends in the same NUTS 3 region were modeled as self-loops.

From the Eurostat database, GDP per capita adjusted for Purchasing Power Parity (PPP) and population data for NUTS 3 regions were utilized as economic background variables. GDP PPP is a widely accepted metric for international territorial

[13]https://www.ksh.hu/regionalatlas_eu_nuts

TABLE 3.1: Applied indicators for ownership network

	v	Indicators	Description	Data source
Node dataset (NUTS 3 regional data)	m_1	TA	Total assets	Amadeus
	m_2	SR	Solvency ratio (Asset based) (%)	Amadeus
	m_3	SH	Shareholders' funds	Amadeus
	m_4	RB	ROE using P/L before tax (%)	Amadeus
	m_5	RCB	ROCE using P/L before tax (%)	Amadeus
	m_6	PM	Profit margin (%)	Amadeus
	m_7	PLF	P/L for period	Amadeus
	m_8	PLB	P/L before tax	Amadeus
	m_9	OR	Operating revenue	Amadeus
	m_{10}	FA	Fixed Assets	Amadeus
	m_{11}	EN	Number of employees	Amadeus
	m_{12}	CR	Current ratio	Amadeus
	m_{13}	CF	Cash flow	Amadeus
	m_{14}	CO	Number of companies	Amadeus
	m_{15}	GDP	GDP/ capita in purchasing power priority	Eurostat
	m_{16}	PI	Patents	PATSTAT
Edges	i	FROM	The NUTS 3 ID of parent companies	Amadeus
	j	TO	The NUTS 3 ID of daughter companies	Amadeus
	$d_{i,j}$	Dist	Distance between regions	Eurostat
	$a_{i,j}$	OWN	Number of ownerships	Amadeus

comparisons Abrham and Vosta (2011), though differences between nominal and PPP GDP within the EU are relatively minor.[14] During the investigation, it was checked whether differences arose from using nominal versus PPP GDP, but no divergence was observed, consistent with findings in (Paas et al., 2008).

GDP data at the NUTS 3 level were unavailable for Iceland (2 regions), Liechtenstein (1 region), Switzerland (25 regions), and the United Kingdom (179 regions); thus, country-level GDP per capita values were applied.[15]

All indicators were aggregated to NUTS 3 regions, with mean values calculated for each region to define metrics $m_1 - m_{14}$.

Two distinct data files were created to facilitate further analysis: one containing node-related information and the other detailing edges between nodes (Table 3.1). The edge file included connections between regions (i and j), inter-regional distances ($d_{i,j}$) sourced from Eurostat, and the variable $a_{i,j}$ quantifying ownership links from region i to j as directed edges.

Table 3.1 summarizes the employed indicators and their sources. Definitions are provided in Appendix A.

3.2.2 Used data for collaboration network

For the collaboration network investigation, a slightly different dataset was used as the main database; the Cordis Horizon 2020 dataset was utilized since the H2020 program relevant information is available in this one. The database that was available for the research contains information about more than 25,000 earned projects and more than 150,000 participants. 86% of the participants came from Europe, as

[14]Source: https://statisticstimes.com/economy/gdp-nominal-vs-gdp-ppp.php
[15]Source: https://ec.europa.eu/eurostat/databrowser/view/NAMA, retrieved: 5 May 2022.

in the projects there were participants from countries outside the EU. During data validation, it was observed that 28% of the participants were individuals and 72% were organizations. Organizations can be different types of entities like universities, colleges, research institutions, governmental entities, public companies, and all other kinds of companies.

This database contains different types of information about the entities like the legal name of the entity, VAT number, address, and different contact information. About projects, it contains the description of the project goals, start and end date of the projects, the earned contribution value (this information is available as a total number for the whole project and in addition the value is also available on entity level, what is the earned amount by the entity from the project budget), funding schema, deliverables, and the sub-project identification number.

In addition to the above-mentioned variables, from the Cordis data source, the Framework Programme 7 (FP7) information was also used to determine whether two entities from the H2020 collaboration had collaborated previously also in FP7 or not. This variable was calculated as a dummy based on any connection in FP7 by the two corresponding entities.

The Cordis database was connected to the Orbis data using the BvD ID number. This identifier was made available for the Cordis dataset by the Bureau van Dijk company based on the entity names, addresses, and geocodes in both datasets. From the Orbis database, the key financial variables were used together with the employee number and the size classification Blažek et al. (2023) (the size classification was calculated and defined also by the owner of the database). The Orbis database contains information in total for more than 100 million entities, so data from Orbis were retrieved for the relevant BvD ID numbers, which are also recorded in the Cordis database as the identification number for any of the participants. The values are available for years, and data for all entities between 2010 and 2019 were used, with the yearly data aggregated to calculate the mean value for all of them.

As mentioned, the main connection field between the datasets is the BvD ID number; the other one is the NUTS code. As in the Eurostat database the information is available at this level, the GDP PPP on NUTS 3 level was integrated, the poverty information for the area on NUTS 2, and the corruption index on NUTS 1 level. The most detailed information available was always used as the main unit was always the NUTS 3 level; larger territorial information was used only in cases where the data was not available for smaller units. This data from Eurostat was used as indicators for the economic background.

From the Patstat database, the number of patent information was used as a proxy indicator to provide information about the technological background of the territories. These data were calculated separately on the NUTS 3 level and integrated into the research database that was used.

Two other indicators were also included in the database, namely the MULTI and the PROG values. These two indicators were calculated on the basis of (Kosztyán et al., 2022a). The value of multi-project membership (MULTI) reflects the number of projects awarded running in parallel at the same time, and the consortium organizations are at least partially common (see Eq. 3.27). The program membership value reflects the projects that were completed earlier—but also within the H2020 program—and have some outcome or result as a mandatory input for the following project(s), so they are dependent on each other (see Eq. 3.28). These membership values are related to projects by default, although the mean value of the organized project can be calculated for each organization. If the mean value of program membership is high for a particular entity, in this case, this high value indicates that this

organization has experience with previously awarded successful projects, which are closely related to the current project the organization is executing. In contrast, a high mean value of the multiproject membership variable for an organization means that this organization organizes several simultaneously running projects. The details of the calculation of these two indicators are discussed in Section 3.3.

After all the above-mentioned data were collected, organized, and connected to each other, 20,172 different organizations were identified for which all the data could be connected.

Two different data files were created, one for the nodes and one for the edges between these nodes. In this work, links were defined in cases where two organizations participated in the same project in H2020. In total, 237,084 edges were identified between the defined nodes. Two versions of the edges dataset were created, one with and one without direction information—the direction here is from the participant to the coordinator. In the end, this edge data file contained 5 variables (see Table 3.2) and 237,084 observations. The node dataset contained the 20,172 different organizations with 59,761 observations. In cases where one organization appeared in different projects, the corresponding variables were aggregated by BvD ID number. This dataset contained in total 201 columns (variables), as for the organizational level data from the Orbis data source were collected for ten years between 2010 and 2019.

In Table 3.2 all the indicators used are collected with a short description, the source of the data, and the level on which the data is available. As mentioned in the table, there are four different variable categories as follows:

1. Corporate indicators include the main profit and loss data together with the balance sheet records of the companies. The main data source is the Orbis database. All indicator values are available on an organizational level.

2. The background indicators of the economy containing GDP / capita, the corruption index (CI) and poverty (Pov). The main source is the Eurostat database. The availability varies between NUTS 1, 2, and 3 levels.

3. The technology indicator used is the number of patents (PI) in NUTS 3 regions whose information comes from the Patstat database from which the aggregation was performed for NUTS 3 regions as a summation.

4. The Collaboration indicators are created separately for nodes and edges, too. Earned contribution (EC) is available directly from the Cordis database on the organizational level as the amount of money that the company receives from the H2020 program. The 'MULTI' and 'PROG' are the two calculated variable as mentioned earlier which reflect to the multiple project involvement and execution within the H2020 umbrella.

   From the edges point of view, two columns have been added to handle the directional information of the connection. 'FROM' stands for the source of the link from organization i, and 'TO' is the destination node j. 'Dist' is the geographical distance between the node i and j that was calculated based on the address information from the Orbis and Cordis databases. The dummy variable FP7 is the added information whether the organization i and j had previously executed any project together within the Seventh Framework Programme. All the edges related information is available on organizational level.

In total 23 different corporate and economic predictors were used together with the nodes and edges-related variables.

TABLE 3.2: Applied indicators for collaboration network

	v	Indicator	Description	Data source	Level	Type
Node dataset	m_1	TA	Total assets	ORBIS	Org.	Corporate
	m_2	SR	Solvency ratio (Asset based) (%)	ORBIS	Org.	
	m_3	SH	Shareholders' funds	ORBIS	Org.	
	m_4	RB	ROE using P/L before tax (%)	ORBIS	Org.	
	m_5	RCB	ROCE using P/L before tax (%)	ORBIS	Org.	
	m_6	PM	Profit margin (%)	ORBIS	Org.	
	m_7	PLF	P/L for period	ORBIS	Org.	
	m_8	PLF	P/L before tax	ORBIS	Org.	
	m_9	OR	Operating revenue	ORBIS	Org.	
	m_{10}	FAs	Fixed Assets	ORBIS	Org.	
	m_{11}	EN	Number of employees	ORBIS	Org.	
	m_{12}	CR	Current ratio	ORBIS	Org.	
	m_{13}	CF	Cash flow	ORBIS	Org.	
	m_{14}	Size	Size of the company	ORBIS	Org.	
	m_{15}	GDP	GDP/capita in purchasing power priority	EUROSTAT	NUTS3	Economy
	m_{16}	CI	Corruptin Index. Scale: 0 (highly corrupt) to 100 (very clean).	EUROSTAT	NUTS1	
	m_{17}	Pov	Poverty	EUROSTAT	NUTS2	
	m_{18}	PI	Patents	PATSTAT	NUTS3	Technology
	m_{19}	EC	Earned Contribution	CORDIS	Org.	Collaboration
	m_{20}	MULTI	Mean of the multiproject membership value	CORDIS	Org.	
	m_{21}	PROG	Mean of the program membership value	CORDIS	Org.	
Edges	i	FROM	The ID of the partner company	CORDIS/ORBIS	Org.	
	j	TO	The ID of the coordinator	CORDIS/ORBIS	Org.	
	D	Dist	Geographical distance between two organizations	CORDIS/ORBIS	Org.	
	FP7		Any earned projects in a collaboration between companies	CORDIS	Org.	

3.3 Methods

In the following Section all the used and employed methods are introduced together with the aim for them and highlighting the benefits and also the limitations for all of them. The method used for which investigation is also described together with the description of them.

3.3.1 Methods applied for ownership network investigation

The hierarchical (directed) link between the parent and subsidiary firms is represented using a binary adjacency matrix $\mathbf{A}$, defined by the following rule:

$$a_{i,j} = \begin{cases} 1 & \text{if the } i\text{-th firm holds ownership over the } j\text{-th firm,} \\ 0 & \text{otherwise} \end{cases} \tag{3.1}$$

Due to the challenges associated with interpreting fractional ownership structures, the actual ownership shares are not incorporated into the model. From now on, the adjacency matrix $\mathbf{A}$ is known as the *Company Ownership Matrix (COM)*. The dataset includes precise spatial coordinates for each firm. However, given that all economic and technological indicators are available only at the NUTS 3 regional level, and in order to preserve company anonymity, the information was aggregated accordingly. Nonetheless, the disaggregated data is retained separately to facilitate link prediction (i.e., the number of ownership connections) between different NUTS 3 regions. Each locality is assigned to its respective NUTS 3 region (county), and the firms are geographically mapped using the incidence matrices $\mathbf{A}^{[\text{mo,NUTS 3}]}$ and $\mathbf{A}^{[\text{da,NUTS 3}]}$, defined as follows:

- $a_{i,j}^{[\text{mo,NUTS 3}]}$ equals one if the i-th parent firm's headquarters is located in the j-th NUTS 3 region,
- $a_{i,j}^{[\text{da,NUTS 3}]}$ equals one if the i-th subsidiary operates within the j-th NUTS 3 region,

Based on these, a directed, weighted network is constructed to capture the number of ownership ties between regions:

$$\mathbf{A}^{[\text{NUTS 3}]} = \left(\mathbf{A}^{[\text{da,NUTS 3}]}\right)^T \times \mathbf{A} \times \mathbf{A}^{[\text{mo,NUTS 3}]}, \tag{3.2}$$

Here, $\mathbf{A}^{[\text{NUTS 3}]}$ denotes the *(aggregated) company ownership matrix (ACOM)*. In cases where both the parent and subsidiary firms are located in the same NUTS 3 region, a self-loop is created at the regional level. The matrix entry $a_{i,j} \in \mathbf{A}^{[\text{NUTS 3}]}$ reflects the number of owners between NUTS 3 region i and region j.

This transformation from the company level to the regional level enables the examination of interregional linkages through cross-sectional analysis conducted on a yearly basis.

When multiple time periods are analyzed, the data structure extends into a three-dimensional array rather than a single adjacency matrix, where the third dimension captures the temporal aspect (i.e., the year). Since the analysis focuses on intercounty relationships, the NUTS 3 notation is dropped for simplicity. The adjacency matrix corresponding to year t is denoted as $\mathbf{A}_t = \mathbf{A}_t^{[\text{NUTS 3}]}$.

Applied Null Models

Null models are applied to predict the likelihood of connections forming between nodes. One of the most commonly used approaches is the random configuration model introduced by Newman and Girvan (2004), which estimates the probabilities of the link by assuming a random network that maintains the original in- and out-degree distributions:

$$a_{i,j} \sim p_{i,j}^{[NG]} = \frac{k_i^{[out]} k_j^{[in]}}{L}, \tag{3.3}$$

where the out-degree $k_i^{[out]} = \sum_j a_{i,j}$, the in-degree $k_j^{[in]} = \sum_i a_{i,j}$, and $L = \sum_i \sum_j a_{i,j}$ denotes the total number of links. The $\sim$ stands for "distributed as" or "generated from" in this equation because the $a_{i,j}$ values are determined from the $p_{i,j}^{[NG]}$ values.

It is important to consider the treatment of self-loops that may emerge as a result of regional-level aggregation of the ownership network. To handle this, Arenas et al. (2008) proposed a multi-resolution approach - known as the AFG method - that adds r self-loops to each node, thereby enhancing node strength without modifying the original structure of the network. This correction is applied by transforming the adjacency matrix into $\mathbf{A}_r = \mathbf{A} + r\mathbf{I}$, where $\mathbf{I}$ is the identity matrix, and r controls the weight of the self-loops for each node.

In graph-theoretical terms, the modified matrix corresponds to the original network augmented with self-loops of weight r uniformly assigned to all nodes. The formulation in Eq. (3.3) involves a uniform additive shift r to the strength of each node. Crucially, this transformation does not alter the intrinsic structure of the network: key metrics such as the strength distribution, weighted clustering coefficients, and arbitrary-order strength correlations remain invariant. This invariance arises because the weights of the existing internode edges, which make up the topological core of the network, are not affected by the addition of self-loops. The shift impacts only the individual node properties in a homogeneous manner across the network. Spectrally, each eigenvalue of the adjacency (or weight) matrix is translated by r, thus preserving all properties that depend solely on eigenvalue differences. In particular, the eigenvectors remain unchanged under this transformation. This adjustment is used in the modularity analysis, though it is known to underestimate the influence of self-loops.

Although the so-called randomized null model defined in Eq. (3.3) is commonly used; it often fails to capture patterns in empirical networks (Liu et al., 2012b). Despite this, it remains a foundation for various community detection methods, including modularity maximization (Newman, 2010b).

One key limitation of this configuration-based model is its disregard for spatial constraints, that is, it does not consider the physical or economic distance between regions. To incorporate spatial effects and attractiveness or importance at the node level instead of the sum of the incoming or outgoing edges, an alternative null model is defined as follows (Barthélemy, 2011; Expert et al., 2011):

$$a_{i,j} \sim p_{i,j}^{[spat]} = \sigma \left(I_i^{[out]}\right)^{\alpha} \left(I_j^{[in]}\right)^{\beta} f(d_{i,j}), \tag{3.4}$$

Here, $I_i^{[out]}$ and $I_j^{[in]}$ represent the importance or attractiveness of the nodes. The importance of a node is characterized by so-called centralities. More details about centralities are discussed in Section 3.3.1. The reason why the $[in]$ and $[out]$ versions of importance are used is the fact that a directed network is applied in this work

and therefore the in- and out-edges need to be handled separately, and therefore the centrality and importance measures need to be applied correspondingly. The α and β are fitting parameters. The normalization constant σ ensures that $\sum_i \sum_j p_{i,j} = \sum_i \sum_j a_{i,j}$ and is computed as

$$\sigma = \frac{L}{\sum_i \sum_j \left(I_i^{[out]}\right)^{\alpha} \left(I_j^{[in]}\right)^{\beta} f(d_{i,j})}.$$

The deterrence function $f(d_{i,j})$ -where $d_{i,j}$ denotes the distance between nodes- is empirically derived by binning techniques such that the prediction error is minimized, following the procedure described in (Expert et al., 2011):

$$f(d) = \frac{\sum_{i,j|d_{i,j}=d} a_{i,j}}{\sum_{i,j|d_{i,j}=d} I_i^{[out]} I_j^{[in]}}. \tag{3.5}$$

Equation (3.4) generalizes the configuration model of Eq. (3.3). When $\alpha = \beta = 1$, $f(d) = 1$, and $\sigma = 1/L$, both models are equivalent. The AFG correction is applicable to distance-dependent prediction as well, but if $f(d_{i,j}) \neq \infty$ for all $d_{i,j} = 0$, then Eq. (3.4) already accounts for self-loops without additional correction. Notably, Eq. (3.4) is a hybrid null model because it blends topological and spatial factors. Unlike Eq. (3.3), which relies solely on degree distributions excluding any other influencing indicator that could determine the weights of the edges between nodes, the spatial model described in 3.4 incorporates distance effects into the model. Moreover, it allows for differential treatment of node importance through its regression parameters, distinguishing the importance of incoming and outgoing edges, which have already different meanings.

To further refine the link predictions, a gravity-based null model can be used, where the deterrence function follows a power law form $f(d_{i,j}) = d_{i,j}^{\delta}$. By adopting the typical formulation of gravity models and substituting node importance variables with m_i (e.g., GDP per capita, population) by Gadár et al. (2018), the null model is generalized as follows:

$$a_{i,j} \sim p_{i,j}^{[grav]} = \tau d_{i,j}^{\delta} \prod_{v=1}^{N} m_{i_v}^{\alpha_v} m_{j_v}^{\beta_v}, \tag{3.6}$$

where N is the number of indicators that belong to the nodes, and α_v, β_v, τ, and δ are the regression coefficients of the model. This formulation of Eq. (3.6) is referred to as the GEN. Provided that $d_{i,j} \neq 0$, the parameters can be estimated via a log-linear transformation of GEN with the aim to be able to use linear regression:

$$\log a_{i,j} \sim \log p_{i,j}^{[grav]} = \log \tau + \delta \log d_{i,j} + \sum_{v=1}^{N} \alpha_v \log m_{i_v} + \sum_{v=1}^{N} \beta_v \log m_{j_v}. \tag{3.7}$$

In my case, all $m_i > 0$, but self-loops imply $d_{i,i} = 0$. To address this, one option is to add a small constant (e.g., 1 km) to all distances so that $\log(d_{i,i} + 1) = 0$ and Eq. (3.7) becomes solvable. However, Burger et al. (2009) noted that this addition of a randomly chosen small number for 0 distance can distort the estimation and recommended applying Poisson regression and solving Eq. 3.6 directly instead of solving Eq. (3.7). To address these findings of Burger et al. (2009) given that the geographic coordinates of the companies are available, the average internal distances within each NUTS 3 region are used to represent the self-loop distances. With this

approach the distortion can be avoided, as no random number is applied without real meaning, but using valid distance values which are relevant for the particular NUTS 3 region. Although this adjustment can be generalized to all region pairs, it had negligible effect and was only implemented for self-loops.

Because Eq. (3.7) is a linear regression model, the standard assumptions - such as normality, homoscedasticity, and independence - must hold. In particular, multicollinearity is checked by using Variance Inflation Factor (VIF):

$$VIF_i = \frac{1}{1 - R_i^2}, \tag{3.8}$$

where $VIF_i \in [1, \infty[$ quantifies the variance inflation factor for variable i, and R_i^2 is the coefficient of the determination of the regression equation:

$$X_i = \alpha_0 + \alpha_1 X_1 + \cdots + \alpha_{i-1} X_{i-1} + \alpha_{i+1} X_{i+1} + \cdots + \alpha_n X_n + \epsilon \tag{3.9}$$

To reduce multicollinearity, the highest VIF value must be less than 2.5. $\max_i VIF_i < 2.5$ (Johnston et al., 2018).

The GEN model Eq. (3.6) and its log-linear form Eq. (3.7) are strictly economic in nature and do not rely on network-based features. Despite this, the hypothesis is that GEN offers better link prediction than other null models. As such, GEN not only enables better edge prediction, but also facilitates the derivation of network metrics like centrality and modularity with reduced prediction error. Nevertheless, it should be emphasized that GEN is purely an economic model, which predicts not only the links, but also the structural characteristics (centralities and modules) of the resulting network.

The goodness of fit of a null model is retrieved from the way how well the edges are estimated in the network. So in case there are variable parameters, the absolute difference between observed and predicted link values must be minimized. Formally:

$$\min \leftarrow \epsilon = ||\mathbf{A} - \mathbf{P}||. \tag{3.10}$$

where $\|A - P\|$ is the Frobenius norm widely used in networks science, machine learning, and numerical linear algebra (Pişcoran, 2021):

$$\|\mathbf{A} - \mathbf{P}\|_F = \sqrt{\sum_{i=1}^{n} \sum_{j=1}^{n} |A_{ij} - P_{ij}|^2} \tag{3.11}$$

This section has introduced three types of null models. The configuration model from Newman and Girvan (2004) is based solely on the network properties during link prediction. The distance-dependent null model by Expert et al. (2011) incorporates both topology and geographic distance, forming a hybrid approach. Although there are several other null models exist Barthélemy (2011), the proposed GEN model is, to my knowledge, the first to predict links using only spatial, economic, technological, and corporate financial indicators without relying on network-property indicators.

Although GEN's optimization minimizes absolute differences in predicted links, its mathematical formulation closely resembles that of gravity-based models, which typically minimize squared errors. This conceptual alignment supports the use of gravity models for predicting both links and therefore the overall ownership network too.

Communities

One of the most prominent uses of null models lies in the detection of communities within networks. Traditional community detection algorithms based on modularity optimization employ metrics $f(C)$, which quantify the difference between the actual number of intracommunity edges and their expected value based on a null model (Newman and Girvan, 2004; Yang and Leskovec, 2015).

$$f(C) = \text{(fraction of edges within communities)} - \text{(expected fraction of such edges)}. \tag{3.12}$$

For the specific case of the directed network under consideration in this study, the difference mentioned above is formulated as follows:

$$f(C) = \frac{1}{L} \sum_i \sum_j \left(a_{i,j} - p_{i,j}\right) \delta\left(C_i, C_j\right), \tag{3.13}$$

where $p_{i,j}$ denotes the predicted number of ownership relations from region i to region j, $a_{i,j}$ is the same value but for the original network, so in other words the actual number of ownership relations from region i to region j and $\delta(C_i, C_j)$ is the Kronecker delta function, which takes the value of one when both nodes i and j belong to the same community, and zero otherwise.

The total modularity of the partition C can then be derived as the cumulative sum of the modularities for each individual community C_c, where $c = 1, \ldots, n_c$ and n_c is the total number of identified communities:

$$M_c = \frac{1}{L} \sum_{(k,l) \in C_c} \left(a_{k,l} - p_{k,l}\right). \tag{3.14}$$

The resulting modularity value M_c associated with a given community C_c may be either positive, negative, or zero. A value of zero indicates that the actual number of links within the community matches exactly the number predicted by the null model. If modularity is positive, it implies that the subgraph C_c exhibits greater internal connectivity than is expected by the null model, thus suggesting the presence of a well-defined community. To identify optimal community structures, the expression in Eq. (3.14) must be maximized.

$$\max \longleftarrow M_c \tag{3.15}$$

So, the edges are included into the communities with which this value is reaching the maximum and getting the optimal community structure. Although Eq. (3.14) is conventionally optimized using the Louvain algorithm Blondel et al. (2008), here the more recent Leiden algorithm Traag et al. (2019) was employed, which offers enhanced stability in the detection of community structures.

For a short explanation of the Leiden algorithm, the main steps are as follows (see also Figure 3.3).

1. *Initialization:* Like the Louvain algorithm, the Leiden algorithm begins with an initial partition of the network where each node forms its own community. This step ensures that the algorithm starts with the maximum granularity possible.

2. *Local optimization:* In this phase, the algorithm iteratively refines the partition. Instead of moving nodes to the community that maximizes the gain in modularity, as done in the Louvain method, the Leiden algorithm employs a more sophisticated approach by moving nodes to arbitrary neighbor communities. This adjustment not only speeds up convergence, but also reduces the likelihood of getting caught in local optima.

3. *Reinforcement of connectivity:* After the local moves, the Leiden algorithm ensures that all communities are connected. This is a notable improvement over the Louvain algorithm, which can yield disconnected communities. The Leiden algorithm guarantees that each detected community is a tightly connected set of nodes.

4. *Aggregation:* Once local optimality and connectivity are ensured for all communities, they are aggregated. The network is then redefined, treating communities as supernodes, and the algorithm iterates from the initialization step with this new reduced graph. This iterative refinement can be executed multiple times, improving the resolution of the community structure.

5. *Termination:* The algorithm concludes when a stable partition is reached, as no further improvements in modularity can be achieved by moving nodes.

When null models based on random configurations are applied, the resulting modules correspond to sets of nodes for which the connections are stronger than expected under random assumptions, meaning that links are more likely to occur within communities than across them (Newman, 2010b). However, in the case of spatial networks, the number of connections between nodes is often influenced by geographic distance (Expert et al., 2011). As a result, the detected modules frequently comprise spatially proximate nodes. This tendency implies that the resulting communities are not only topologically meaningful but also may reflect real-world spatial clustering. If these modules coincide with larger regional units, such as national boundaries or other administrative regions, then it becomes plausible to infer the influence of additional structural or institutional forces in shaping these patterns. Consequently, a key question arises as to whether the modules correspond to larger geographic regions.

Distance-dependent community structures inherently account for the spatial separation between regions. As a result, the detected modules can be interpreted as communities that are independent of regional distances. In other words, this allows for the examination of how the network would behave in the absence of geographic constraints.

In the context of gravity-based models, such modules can be interpreted as representing the *area of investments* Gadár et al. (2018), and are henceforth referred to as *EICs*. These EICs represent a collection of regions where the intensity of investment relationships, quantified by the number of ownership links, exceeds the levels expected solely based on economic, financial, technological indicators and spatial proximity. If the boundaries of these EICs coincide with existing administrative borders, it suggests that such administrative divisions act as a primary structural force shaping investment patterns, which carries implications for policy making at the European Union level.

This study introduces a generalization of gravity-based null models (GEN), designed to identify and interpret the factors, both economic and technological, that drive the emergence of regionally concentrated investment zones.

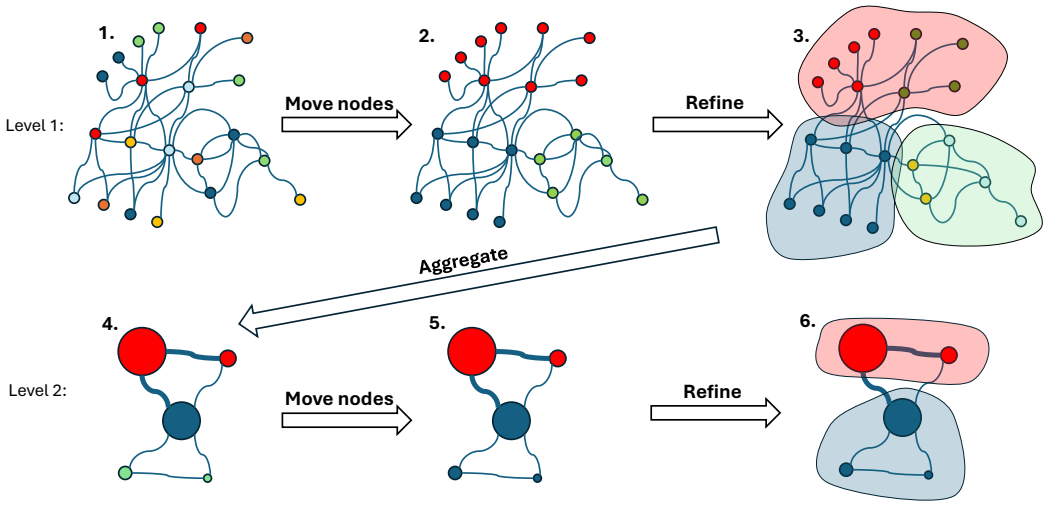


FIGURE 3.3: The Leiden algorithm initiates with a singleton partition (1). It proceeds by relocating individual nodes between communities to identify a partition (2), which is subsequently refined (3). An aggregate network (4) is then constructed based on the refined partition, utilizing the non-refined partition to establish an initial partition for the aggregate network. For instance, the red community in (2) is refined into two subcommunities in (3), which, following aggregation, correspond to two distinct nodes in (4), both assigned to the same community. The algorithm then repositions individual nodes within the aggregate network (5). In this instance, the refinement does not alter the partition (6). These steps are iteratively performed until no further improvements are achievable. Source: recreated based on Traag et al. (2019)

Given that a company ownership network Company Ownership Network (CON) typically represents a static snapshot of ownership relations, dynamic analyzes require the extension of the model over time. One way to achieve this is by constructing a multilayer network in which each layer corresponds to a specific year. In such a formulation, the null models operate independently on each layer, allowing simultaneous predictions over time. Alternatively, the network can be conceptualized as a dynamic structure, wherein links are defined over continuous temporal intervals. Although dynamic models are more suitable for continuous-time data, multilayer networks - composed of discrete, yearly static snapshots - are more compatible with the structure of existing null models, facilitating their extension into temporal analysis.

Both Louvain and Leiden algorithms can be extended to multilayer network structures, in which each layer corresponds to a time slice. Consequently, the proposed gravity-based null models can be used for link prediction in such multilayer networks, enabling the identification of *yearly EIC*.

Through the application of link prediction methods, it becomes possible to model the entire evolution of the network. This approach not only enables the estimation of

various network properties, such as centrality measures, but also provides a framework for interpreting their formation mechanisms and predicting their temporal development.

Multilayer Network as a Discrete Representation of a Spatial-Temporal Network

A multilayer network is formally defined as a pair $\mathcal{M} = (\mathcal{G}, \mathcal{C})$, where $\mathcal{G} = \{G_\alpha = (V_\alpha, E_\alpha, W_\alpha), \alpha \in \{1, \ldots, m\}\}$ represents a collection of weighted graphs (directed or undirected), which are also called layers of $\mathcal{M}$. Each layer G_α consists of a set of vertex V_α (the set of nodes), an edge set $E_\alpha \subseteq V_\alpha \times V_\alpha$ (i.e. the set of links or arcs), and a weight function $W_\alpha : V_\alpha \times V_\alpha \to \mathbb{R}_0^+$ assigning weights to the edges within layer α of graph G_α.

The set of inter-layer connections is defined as

$$\mathcal{C} = \{E_{\alpha,\beta} \subseteq V_\alpha \times V_\beta, W_{\alpha,\beta} : V_\alpha \times V_\beta \to \mathbb{R}_0^+, \alpha, \beta \in \{1, \ldots, m\}, \alpha \neq \beta\}, \tag{3.16}$$

which captures the linkages between nodes residing in different layers $G_\alpha, G_\beta \in \mathcal{M}$, with the condition that $\alpha \neq \beta$.

In current work, inter-layer connections are not explicitly modeled; hence, it is assumed that $\mathcal{C} = \emptyset$. In the context of spatial-temporal networks, each layer may correspond to a distinct temporal instance (e.g., a specific year), thus denoted as $\alpha = t$. Since the geographical regions remain fixed over time, the node set is identical across all time slices, i.e., $V_t = V, \forall t$. The temporal dynamics are reflected solely in the edge weights, which may vary annually. Consequently, regional connections can be estimated independently for each year using a yearly gravity-based model, as shown in Eq. (3.17):

$$\log a_{i,j,t} \sim \log p_{i,j,t}^{[grav]} = \log \gamma_t + \delta_t \log d_{i,j} + \sum_{v=1}^{N} \alpha_{v_t} \log m_{i_{t,v}} + \sum_{v=1}^{N} \beta_{v_t} \log m_{j_{t,v}}. \tag{3.17}$$

The temporal variation in the estimated regression parameters reflects changes in the influence of geographic, economic, and technological factors. Furthermore, evaluating embeddedness through multilayer centrality metrics provides insight into the shifting roles of individual regions within the investment network.

Lastly, an examination of how community structures evolve in both space and time can reveal changes in the composition of economic-investment communities (EICs). At the same time, identifying communities within the multilayer framework offers a view of time-invariant EICs, representing stable regional investment patterns throughout the temporal dimension.

Centralities

In network science, centralities are traditionally employed as descriptive metrics to identify key nodes within a network. However, when not only individual links but the entire network structure can be predicted, centralities can subsequently be computed for the predicted network. In this way, centralities themselves also become the subject of prediction. This predictive capability enables researchers to examine which indicators contribute to a region acquiring a central or influential position within the network. To perform such analysis effectively, it is essential that centralities are modeled with as much accuracy as possible.

Given that the network in question is directed — capturing the hierarchical structure of mother-daughter company relationships — only directed and appropriately generalized centrality measures are applied. These include in-degree, out-degree, betweenness, in-closeness, out-closeness, authority, hub, and PageRank centralities.

Degree centrality refers to the number of connections that a node maintains. In directed networks, this concept is divided into in-degree (the number of incoming edges) and out-degree (the number of outgoing edges). The formal definitions of these metrics for a vertex v are the following.

$$C_D v = k_v, \tag{3.18}$$
$$C_D^+ v = k_v^{[in]}, \tag{3.19}$$
$$C_D^- v = k_v^{[out]}. \tag{3.20}$$

Where $k_v^{[in]}$ and $k_v^{[in]}$ were defined at Eq. 3.3, and $k_v = \sum_{j=1}^{N} a_{vj}$. The main reason to introduce these new notation is to be able to refer to them easier in the Results section.

In connected networks, *closeness centrality* (or closeness) captures how near a node is to all other nodes, based on the length of the shortest paths. A node with higher closeness centrality is, on average, closer to all other nodes in the graph.

The formal definition, introduced by Bavelas (1950), defines closeness as the reciprocal of farness:

$$C_c = \frac{1}{\sum_w d(v,w)}, \tag{3.21}$$

where $d(v,w)$ denotes the shortest path length (graph distance) between nodes v and w. In undirected graphs, directionality is irrelevant, whereas in directed graphs, distinguishing between outgoing and incoming distances leads to significantly different outcomes.

Betweenness centrality (C_B) quantifies the extent to which a node lies on the shortest paths between other pairs of nodes. Nodes with high betweenness centrality frequently act as intermediaries or bridges within the network, indicating their potential influence in information or resource flow.

$$C_{Bj} = \sum_{i \neq j \neq k} \frac{n_{ik}^{\mathrm{SP}}(j)}{n_{ik}^{\mathrm{SP}}} \tag{3.22}$$

where n_{ik}^{SP} is the total number of Shortest Paths (SPs) from node i to node k and $n_{ik}^{\mathrm{SP}}(j)$ is the number of paths that pass through node j.

PageRank centrality is determined by the following system of equations which can be solved recursively:

$$v_i = \alpha \sum_j a_{ji} \frac{v_j}{C_D j} + \frac{1-\alpha}{N}, \tag{3.23}$$

where $C_D j$ is the Degree centrality of node j, α is a damping factor in the range $[0,1]$, $i,j \in V$, $\sum_i v_i = 1$ and N represents the total number of nodes in the network.

Hub centrality (C_H) and authority centrality (C_A) are also computed to rank nodes according to their structural role. Hub scores capture a node's tendency to link to well-connected nodes, while authority scores measure how often a node is the target of links from other nodes.

This concept was formalized by Kleinberg (1999) through the development of the hyperlink-induced topic search (HITS) algorithm. The mathematical definitions that encapsulate this intuition are as follows:

Authority centrality: a node that is pointed to by many hubs, meaning it receives links from nodes with high hub centrality.

Hub centrality: a node that points to many authorities, i.e., nodes with high authority centrality.

The corresponding mathematical formulation is as follows.

$$C_A = \alpha A y \tag{3.24}$$

$$C_H = \beta A^T x \tag{3.25}$$

where C_A and C_H represent the vectors of authority and hub centralities, respectively. Here, A is the adjacency matrix of the directed network and A^T denotes its transpose.

Together, these two equations imply that the authority and hub centralities are the eigenvectors of the matrices AA^T and A^TA, respectively, associated with the same eigenvalue. This eigenvalue must be the largest (leading) one, following arguments similar to those used in eigenvector-based centralities. The scalar parameters α and β are free scaling factors; without loss of generality, either α or β can be set to 1, since we are typically interested in the relative, rather than the absolute, centrality values.

The link prediction models proposed by Expert et al. (2011) and Newman and Girvan (2004), and the generalized gravity-based null model (GEN) introduced in this work allow reconstruction of predicted network structures. As such, centrality measures can be derived for both observed and predicted versions of the network. The accuracy of centrality prediction is evaluated by computing the average absolute error:

$$\epsilon_C = \frac{1}{N} \sum_v |C(v) - \widehat{C}(v)|, \tag{3.26}$$

where $C(v)$ is the original centrality value of node v, $\widehat{C}(v)$ is its predicted counterpart, and N is the total number of nodes in the network.

It is important to emphasize that in cases where ϵ_C is small, the GEN - based prediction — which relies exclusively on economic, corporate financial, and technological variables — can be considered as indirectly modeling centrality values. This indicates that the GEN model effectively captures the structural prominence of regions in the network by modeling the combinations of spatial, economic, financial and technological indicators that contribute to their role-player status.

3.3.2 Methods applied for collaboration network investigation

From the database explained in Section 3.2.2, one training sample and one test sample were created. It is important to emphasize that in databases designed for link prediction tasks, the data are typically highly imbalanced, as two organizations are much more likely to have no collaborative relationship than to be involved in a joint project. If this imbalance is not adequately addressed, most machine learning methods are prone to produce biased results.

To address this issue, pairs of organizations with existing collaborative relationships were first selected and divided into the training and test datasets, applying

splitting ratios of 0.7 for the training set and 0.3 for the test set. Subsequently, p pairs of organizations without any recorded collaboration were randomly selected and added to balance the datasets. After preparing these samples, the parameterized models were evaluated in the complete database.

The value of p was optimized by selecting the configuration that resulted in the best accuracy and F1 scores (see details in Section 3.3.2) across $N = 100$ runs for most of the applied methods throughout the entire project. Following the tuning of the p parameter, the highest F1 scores were obtained with $p = 1.302585$.

When assuming no links between any organizations, the baseline accuracy (referred to as Accuracy Null) is 0.767704. Therefore, all link prediction methods are expected to achieve significantly higher accuracy than this baseline value.

Computation of Coefficients from the Precedence Structure

Distance matrices between projects are calculated based on five distinct dimensions: temporal overlap (d_t), precedence relations (d_p), ownership similarity (d_o), project volume (d_v) and textual description (d_{xy}). For an in-depth explanation of how these distance metrics are defined and calculated, refer to (Kosztyán et al., 2022a). Using the resulting pairwise distance values, it is possible to derive the multiproject (MULTI) and program membership (PROG) coefficients for each project as follows:

$$\mathcal{M}_m(p_i) = \max_j \left(1 - d_t(p_i, p_j)\right) \cdot \left(1 - d_o(p_i, p_j) p_i, p_j)\right) \tag{3.27}$$

$$\mathcal{M}_p(p_i) = \max_j \left(1 - d_p(p_i, p_j)\right) \cdot \left(1 - d_{xy}(p_i, p_j)\right) \tag{3.28}$$

In these expressions, p_i and p_j refer to the i^{th} and j^{th} projects, respectively, within the H2020. The terms $\mathcal{M}_m(p_i)$ and $\mathcal{M}_p(p_i)$ represent the computed multiproject and program membership values associated with project p_i, respectively.

Applied Machine Learning Methods for Link Prediction

Accuracy metrics are essential for evaluating the performance of classification models. One of the standard tools for this purpose is the confusion matrix, which provides a comparison between the model's predicted outcomes and the actual observed results.

In this work, a range of standard performance indicators were calculated: *accuracy* (the ratio of correct predictions to the total number of cases), *precision* (measures the proportion of predicted positive cases that are actually positive), *sensitivity* (also known as recall, representing the ratio of correctly predicted positive cases to all actual positives), *specificity* (also known as true negative rate, the ratio of correctly predicted negatives to the total number of actual negatives), *prevalence* (the proportion of actual positives within the population), and the *F1 score*, which is the harmonic mean of precision and sensitivity.

These indicators are widely used in the evaluation of classification tasks, and the choice of which measure to prioritize depends on the specific context and objectives of the analysis. In the case of the H2020 collaboration network research, the focus was on predicting collaboration links between organizations. While accuracy and F1 score were the primary evaluation metrics, all other performance measures were also reported for completeness.

The Link Prediction Problem The collaboration network between organizations was defined as a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V}$ is the set of nodes representing the organizations, and $\mathcal{E}$ is the set of directed edges indicating collaborative links from a partner to a coordinator organization. Each edge in this network is binary: if $e_{i,j} \in \mathcal{E}$ and $v_i, v_j \in \mathcal{V}$, it implies that the organization i has collaborated with the organization j as a partner in at least one joint project.

All machine learning models applied in this study were formulated to solve the following link prediction task:

$$e_{i,j} \sim \hat{y}_{i,j} = f(x_{i,j}) \tag{3.29}$$

In this formulation, $x_{i,j} = [m_{i,k}, m_{j,k}, D_{i,j}, FP7_{i,j}]$ represents the feature vector associated with the node pair (i, j). The output $\hat{y}_{i,j}$ denotes the predicted presence or absence of a link between organizations i and j, and f refers to the machine learning prediction function.

The term $m_{i,k}$ represents the value of the indicator $k \in \{1, 2, \ldots, 19\}$ for the organization i, and $D_{i,j}$ is the geographic distance between organizations i and j.

The binary indicator $FP7_{i,j}$ is defined as:

$$FP7_{i,j} = \begin{cases} 1 & \text{if organizations } i \text{ and } j \text{ collaborated in the FP7 framework,} \\ 0 & \text{otherwise.} \end{cases}$$

Generic Machine Learning Methods Binary *LogR* is a widely used statistical approach in machine learning that addresses binary classification problems by modeling the relationship between a set of input variables and a binary outcome. It outputs a probability that a given instance belongs to the positive class (e.g., class label 1), conditioned on the input features.

In the context of link prediction in collaboration networks, binary LogR is applied to estimate whether a link (i.e., a collaborative relationship) exists between two nodes (i.e., organizations) based on their attributes (m_1 through m_{19}) and the structural characteristics of the network, such as geographic distance $D_{i,j}$. The dependent variable is binary, representing the existence or absence of a link between two organizations.

Binary LogR can be a powerful method for link prediction in collaboration networksSong et al. (2022), as it is capable of modeling complex nonlinear associations between input variables and the target outcome and performs well in high-dimensional feature spaces. Moreover, it offers a probabilistic output value that can be used to rank candidate links by the likelihood of their formation.

Nonetheless, like any machine learning method, binary LogR has its own set of assumptions and limitations. Specifically, it assumes a linear relationship between the input and target variables on the log-odds scale. If this assumption is violated — such as in the case of highly nonlinear relationships or sparsely populated feature spaces — its performance may decline.

Discriminant analysis (DA), such as binary *LDA*, is another statistical technique frequently used in binary classification tasks. It seeks a linear combination of input features that best separates the two classes.

The binary LDA operates by first estimating the means and variances of each feature for both classes. Then it calculates a linear discriminant function designed to maximize the separation between these classes. This function is constructed as a

linear combination of the input variables, where the objective is to maximize the distance between the class means while minimizing the variance within each class. The resulting values of this function are used to determine a decision boundary that separates the classes. In the context of link prediction in collaboration networks, binary LDA can help identify which organizational and structural features most strongly influence the likelihood of a link forming.

Binary LDA offers several advantages in link prediction applications, including its ability to handle high-dimensional data, interpretability, simplicity, and the ability to reflect the underlying data structure (Gu et al., 2011). However, it is also based on certain assumptions. LDA assumes a linear relationship between the independent variables and the dependent class membership. It also presumes that the independent variables follow a normal distribution within each class and that these variables have equal variance across the groups. Violations of these assumptions can introduce bias and reduce predictive accuracy. Furthermore, LDA can be sensitive to multicollinearity - when independent variables are highly correlated — which can lead to unstable estimates of the discriminant function.

Both linear regression and linear discriminant analysis aim to estimate the following equation, which serves as a linear approximation of Eq. (3.29):

$$e_{i,j} \sim \beta_0 + \beta_D D_{i,j} + \beta_{FP7} FP7_{i,j} + \sum_{k=1}^{19} (\beta_{i,k} m_{i,k} + \beta_{j,k} m_{j,k}) \tag{3.30}$$

In some cases, LDA can be replaced with *QDA* to improve predictive performance (Tharwat, 2016). The principal advantage of using QDA over LDA is its ability to model nonlinear relationships between the input variables and the target class label (Tharwat, 2016). Aside from relaxing the linearity assumption, QDA shares the same general assumptions as LDA - including normality of independent variables within each class and equality of class variances - and may be similarly affected by violations of these assumptions.

Non-generic Machine Learning Methods To predict collaborative links between organizations, the *Support Vector Machine (SVM)* algorithm was used, as introduced in Hearst et al. (1998), and applied in similar contexts by (Yu et al., 2020). SVM is a discriminative classification method that constructs a decision boundary to separate data points into distinct classes. However, it does not yield class membership probabilities. For each organizational pair $e_{i,j}$, a feature vector $x_{i,j}$ was constructed and the model SVM was applied to produce a predicted label $\hat{y}$. When $\hat{y} = 1$, the model indicates a high probability of collaboration, while $\hat{y} = -1$ implies a low likelihood of collaboration.

The performance and flexibility of the SVM model are governed by several key hyperparameters (Gold and Sollich, 2003):

- *C*: Controls the trade-off between maximizing the decision margin and minimizing classification error.
- *Kernel type*: Determines the transformation function applied to map the data into a higher-dimensional space. Common choices include linear, polynomial, Radial basis function (RBF), and sigmoid kernels.
- *Kernel coefficient*: Influences the shape and flexibility of certain kernels.
- *Degree*: Specifies the degree of the polynomial kernel.

- *Gamma*: Defines the width of the Gaussian function used in the RBF kernel.
- *Class weights*: Adjusts the model to compensate for class imbalance by assigning different importance levels to the classes.

Despite its strengths, SVM has several limitations. The method is not well-suited to large-scale datasets due to its computational cost, and its performance is sensitive to the selection of the kernel function. It is also susceptible to noise in the input data and generates a binary classifier that is difficult to interpret. Furthermore, the model's decision function coefficients do not provide intuitive insights into feature importance.

In addition, the *RF* algorithm was also applied as described in (Pal, 2005). RF is an ensemble-based approach that constructs multiple decision trees using randomly selected subsets of the data and features. Each tree produces an individual prediction, and the final output is determined by majority voting across all trees. This method is capable of modeling complex, nonlinear relationships and performs effectively in high-dimensional spaces. Moreover, RF can be used to assess the relative importance of features, which aids in understanding the underlying mechanisms driving collaboration within the network.

The prediction function of the RF model, applied to a pair of organizations $(e_{i,j} \sim \hat{y}_{i,j})$, can be expressed as follows:

$$e_{i,j} \sim \hat{y}_{i,j} = f(x_{i,j}) = \text{mode}([T_1(x_{i,j}), T_2(x_{i,j}), \ldots, T_M(x_{i,j})]) \tag{3.31}$$

In this formulation, f represents the prediction function, $x_{i,j}$ denotes the feature vector that describes the relationship between organizations i and j, and $T_1, T_2, \ldots, T_M$ are the individual decision trees that constitute the ensemble.

The primary hyperparameters of the RF model, as identified by Probst et al. (2019), include the following:

- *n_estimators*: Specifies the total number of trees in the forest. Increasing this number can enhance prediction performance, but also raises computational costs.
- *max_depth*: Defines the maximum allowable depth of each decision tree. Although deeper trees can capture more complex patterns, they also increase the risk of overfitting.
- *min_samples_split*: Sets the minimum number of samples required to split an internal node. Higher values can help prevent overfitting by enforcing stricter splitting conditions.
- *min_samples_leaf*: Determines the minimum number of samples needed to form a leaf node. This parameter also contributes to the control of overfitting.
- *max_features*: Indicates the number of features considered in determining the best split. Using options such as "sqrt" or "$\log_2$" helps introduce randomness and can mitigate overfitting by reducing the correlation among trees.
- *bootstrap*: Specifies whether bootstrapping is applied when sampling the data for each tree. When set to "true," each tree is trained on a randomly selected subset of the dataset, which promotes diversity among the trees and reduces their correlation.

LogR is employed to identify significant predictors in the link prediction task. However, it assumes that the relationship between variables is linear. Consequently, this method becomes unsuitable when the associations between variables are non-linear or when the classes are not linearly separable (Tonkin et al., 2012).

To overcome this limitation, the random forest-based method — *Boruta* — was used, as an alternative to the significance analysis that is usually performed with LogR. Boruta is a feature selection algorithm introduced in Kursa et al. (2010), designed to determine which features are truly relevant by comparing their importance with that of shadow features. Shadow features are constructed by permuting the values of each original feature, effectively destroying their association with the target variable. Boruta then employs a random forest classifier to assess the importance of both original and shadow features, based on the average decrease in impurity observed across the ensemble.

Features whose importance exceeds that of their corresponding shadow feature are retained as relevant, while those that do not show significantly higher importance are eliminated. By making this comparison, Boruta effectively identifies features that are genuinely informative, distinguishing them from those that may appear important due to random chance. This selection process helps reduce model complexity, mitigates overfitting, and contributes to improved predictive performance.

RF itself is a widely used ensemble learning method for classification tasks (Sagi and Rokach, 2018). Despite its effectiveness, it exhibits several limitations (Ahmad et al., 2018). Primarily, it functions as a black-box model, making it difficult to interpret the internal decision-making process. This lack of transparency can be problematic in applications where interpretability is critical, such as in regulatory or ethical settings. Additionally, RF can be computationally expensive and memory-intensive, especially when applied to large-scale datasets or when a high number of trees are used in the ensemble.

The performance of RF depends on a balance between bias and variance. While increasing the number of trees can reduce variance and enhance generalization, it may also lead to increased training time and memory usage. Moreover, RF can struggle with highly correlated features Zhu (2020), which may introduce redundancy and confusion during the splitting process. Another challenge is its limited effectiveness on imbalanced datasets, where it may favor the majority class and perform poorly on minority instances.

Extreme Gradient Boosting (XGBoost) Sheridan et al. (2016) is another advanced machine learning algorithm suited for supervised learning tasks, including both classification and regression. Its popularity comes from its high prediction accuracy and computational efficiency.

Once the features are defined, historical data can be used to train the XGBoost model for predicting future collaborations between pairs of organizations. The trained model can then be used to identify potential future partnerships. XGBoost employs decision trees as base learners, and these trees are trained sequentially, where each new tree aims to correct the errors made by its predecessors.

To further improve performance and prevent overfitting, XGBoost includes a regularization term in its objective function. It also applies gradient-boosting techniques Natekin and Knoll (2013), using gradients from the loss function to iteratively update the model parameters in a direction that minimizes the prediction error. Several hyperparameters — including the number of trees, the learning rate, and tree depth — can be adjusted to fine-tune the model's behavior.

TABLE 3.3: Summary of the advantages and limitations of machine learning methods applied to link prediction tasks

Method	Advantages	Limitations
LogR	Intuitive and easy to interpret	Restricted to linear relationships between variables
LDA	Effectively distinguishes between multiple classes	Assumes normally distributed features and equal class covariances
SVM	Performs well with high-dimensional feature spaces	Highly sensitive to the selected kernel function
RF	Capable of modeling non-linear feature interactions	May overfit when exposed to noisy or redundant data
XGBoost	Offers high predictive performance and robustness	Requires extensive preprocessing and careful hyperparameter tuning

Despite its strengths, XGBoost also has notable limitations (Demir and Şahin, 2022). The model is prone to overfitting, particularly if its many hyperparameters are not carefully tuned. The tuning process itself can be challenging due to the large number of adjustable parameters. Additionally, XGBoost has limited native support for categorical variables, often requiring one-hot encoding, which can significantly expand the feature space and increase the risk of overfitting.

From a computational standpoint, training XGBoost on large datasets can be resource-intensive and time-consuming. The algorithm also requires careful feature engineering, as irrelevant or poorly constructed features can adversely affect its performance. Furthermore, XGBoost functions as a black-box model, making its internal logic difficult to interpret and thus less suitable for applications requiring explainability.

Although XGBoost and RF share several common hyperparameters, such as the number of trees (`n_estimators`) and the maximum tree depth (`max_depth`) — XGBoost also includes additional parameters. These include:

- *subsample*: the fraction of observations randomly selected for each tree,
- *colsample_bytree*: the proportion of features randomly selected per tree,
- *gamma*: the minimum loss reduction required to make a further split,
- *reg_lambda*: an L_2 regularization term used to reduce overfitting.

Together, these parameters help to control the complexity of the model and improve generalization, especially in high-dimensional feature spaces.

The main advantages and limitations mentioned above are collected for the methods in Table 3.3.

Parameter tuning of the machine learning methods Tuning hyperparameters is essential to maximize the performance of machine learning algorithms for a specific task or data set. Several well-established techniques are commonly used for this purpose Alibrahim and Ludwig (2021), including:

- *grid search*

- *random search*
- *Bayesian optimization*

Grid search is a widely adopted approach in hyperparameter optimization, where a predefined set of values is specified for each hyperparameter. The algorithm then exhaustively evaluates all possible combinations of these values using a performance metric — often based on cross-validation. The configuration that yields the best evaluation score is selected as the optimal set of hyperparameters. Although grid search is thorough and systematic, it can become computationally demanding when applied to models with many hyperparameters or large search spaces.

In contrast, *random search* Bergstra and Bengio (2012) randomly samples hyperparameter combinations from defined distributions or value ranges. A model is trained and evaluated for each sampled combination, and this process continues for a fixed number of iterations or until a satisfactory performance is achieved. Although it does not guarantee finding the global optimum, random search often identifies high-performing configurations more efficiently than grid search, particularly when only a subset of hyperparameters significantly affects model performance (Mantovani et al., 2015).

Bayesian optimization Turner et al. (2021) adopts a probabilistic approach by iteratively building a surrogate model of the objective function to guide the search for optimal hyperparameters. Typically implemented using Gaussian processes or other probabilistic regression techniques, this method estimates both the value of the function and its uncertainty. At each iteration, the surrogate model is updated based on the observed performance of the hyperparameters previously tested. An acquisition function — such as expected improvement — is then used to determine the next point to sample. This function balances the trade-off between exploring regions of the hyperparameter space with high uncertainty and exploiting regions known to yield good results. The process repeats until a satisfactory solution is found or a predefined number of iterations are reached. Bayesian optimization is particularly well-suited for high-dimensional or non-convex search spaces, where it efficiently navigates toward promising regions by leveraging its probabilistic model.

In this study, the Bayesian optimization framework was used following the methodology described in (Shahriari et al., 2015). The procedure was initialized with a set of starting hyperparameter configurations (see Table 3.4). A Gaussian process was employed as the surrogate model, fitted using the initially evaluated hyperparameters and their associated performance metrics. The expected improvement acquisition function was used to propose subsequent hyperparameter sets, balancing the dual objectives of exploration (sampling in less-certain regions) and exploitation (sampling where high performance is likely).

Each selected configuration was evaluated using cross-validation and the resulting performance metrics were fed back into the surrogate model, allowing it to update its internal representation of the objective function. This cycle of model update and hyperparameter selection was repeated for a predefined number of iterations or until the convergence criteria were satisfied. The goal of the procedure was to identify the hyperparameter set that maximized the performance of the machine learning model.

TABLE 3.4: The hyperparameters and the search ranges applied during tuning of the machine learning methods.

Algorithm	Hyperparameter	Values
SVM	Classification machine	C, μ, One
	kernel	sigmoid, radial
	kernel parameters (γ)	$2^{-5}, 2^{-4}, \ldots, 2^{5}$
	cost of constraints violations	$2^{-5}, 2^{-4}, \ldots, 2^{5}$
RF	Number of estimators	20, 30, 40, 50, 75, 100, 120
	Split criterion	gini, entropy
	Max depth of trees	10, 12, 14, 16, 18, 20, 25
	Minimum number of samples required to split an internal node	2, 3, 4
	Minimum number of samples required to be at a leaf node	1, 2, 3, 4
XGBoost	Number of estimators	20, 30, 40, 50, 75, 100, 120
	Max depth of trees	10, 12, 14, 16, 18, 20, 25
	Learning rate (η)	[0.01,0.5]
	Subsample ratio of columns when constructing trees	0.25, 0.5, 0.75, 1
	Subsample ratio of the training instance	0.25, 0.5, 0.75, 1
	L2 regularization (λ)	[-1,1]

Gini Index calculation

The Gini index by Pérez et al. (2015) is a key metric to measure the impurity of the nodes in decision trees and the importance of variables in random forests. Its calculation and application in variable importance analysis are as follows.

$$\text{Gini} = 1 - \sum_{i=1}^{C} p_i^2 \tag{3.32}$$

Where C is the number of classes and p_i is the proportion of samples in class i. The lower value of the index indicates purer nodes (a Gini index of 0 means that all observations belong to one class).
The reduction in Gini impurity (ΔGini after splitting a node is:

$$\Delta\text{Gini} = \text{Gini}_{\text{parent}} - \left(\frac{N_{\text{left}}}{N_{\text{parent}}} \cdot \text{Gini}_{\text{left}} + \frac{N_{\text{right}}}{N_{\text{parent}}} \cdot \text{Gini}_{\text{right}} \right) \tag{3.33}$$

where $\text{Gini}_{\text{parent}}$ is the Gini index of the parent node, $\text{Gini}_{\text{left}}$ and $\text{Gini}_{\text{right}}$ are the Gini indices of the left and right child nodes, N_{parent}, N_{left}, N_{right} are the number of samples in the parent, left child, and right child nodes, respectively.
Total importance of feature X_j in a random forest:

$$\text{Importance}(X_j) = \sum_{t=1}^{T} \sum_{s \in S_j^t} \Delta\text{Gini}_s \tag{3.34}$$

where T is the total number of trees and S_j is the set of splits in X_j in tree t. Therefore, s denotes one particular split within S_j.

Analyzed Network Properties

The calculation and comparison of network indicators between the original and predicted collaboration networks serve three primary purposes:

- *First*, evaluating network structure indicators reveals how effectively the link prediction methods have succeeded in replicating the structure of the original network.
- *Second*, in the case of non-black-box prediction models, it becomes possible to obtain not only a model to predict collaborations but also models for predicting derived network properties (Kosztyán et al., 2022b).
- *Third*, in many studies, community detection is based on configuration models Girvan and Newman (2002), under the assumption that organizations within a collaboration network are connected randomly. In such cases, communities are considered to form in areas where the densities of internal links are higher than the densities between different communities (modules) (Newman and Girvan, 2004). However, when an edge prediction model is applied as a null model Gadar et al. (2018), it allows the identification of communities where internal link densities exceed not only the random expectation but also the predictions of the model itself (Kosztyán et al., 2021).

To assess the quality of the predictions, there were utilized both node-level indicators (such as various centrality measures) and network-level or structural indicators (such as centralizations, assortativities, and modularity values) to compare the original and the predicted networks. The comparison of node-level indicators reflects not only the alterations in individual link structures but also changes in the nodes' embeddedness. Meanwhile, the comparison of network-level indicators provides insight into the structural differences between the two networks.

The following node-level indicators were calculated to validate the predicted links:

- *DC*: The number of direct connections (degree) of a node divided by the maximum possible number of direct connections is used to identify influential nodes with numerous collaborations (mentioned earlier, see 3.18).
- *Betweenness centrality (BC)*: The proportion of shortest paths that pass through a given node, highlighting nodes that act as intermediaries within the network. (mentioned earlier, see 3.22)
- *Eigenvector centrality (EVC)*: The sum of the centrality scores of a node's neighbors, identifying nodes that are connected to other prominent nodes.
- *Centrality*: A general measure representing the relative importance or influence of a node or an edge, based on degree, positional, or relational characteristics within the network.

EVC quantifies the significance of a node within a network by considering the importance of its neighboring nodes. It is determined by summing the centrality scores of all nodes directly connected to the node in question. This measure is particularly valuable for identifying nodes that are not only well connected but are also linked to other highly influential nodes within the network.

$$\mathrm{EVC}_i = \frac{1}{\lambda} \sum_{i \neq j} g_{ji} \mathrm{EVC}_j \tag{3.35}$$

where λ is some scalar $\lambda > 0$, $i, j \in N$, $N = \{1, \ldots, n\}$ is the set of nodes, $g_{ij} = [g_{ij}]_{i,j \in N}$ where $g_{ij} = 1$ indicates a link from i to j and $g_{ij} = 0$ indicates that such a link does not exist. As the formula shows, the centrality of each node i is proportional to the sum of the centrality of its neighbors. Note that g_{ji} stands for directed edges, as this centrality relies on the link which is pointed toward node i, which can also be applied as is for undirected links.

In addition to node-level metrics, network-level indicators were also evaluated:

- *Modularity*: Measures the extent to which the actual number of links within communities deviates from the expected number under a random connection model, identifying clusters of nodes with stronger internal collaboration. In Section 3.3.1 all the details was mentioned already, the application for collaboration network is very similar. The main difference is that the directed edges do not target the subsidiary of the entity but the coordinator of the project from the partner organizations. The Eq. 3.13 and Eq. 3.14 can also be used here.

- *Assortativity*: Assortativity quantifies the tendency of the nodes within a network to connect to other nodes that share similar attributes or characteristics. It is calculated as the correlation between the degrees of pairs of connected nodes. This metric is particularly useful for detecting collaboration patterns between nodes with comparable features, such as organizations of similar size or profile.

- *Centralizations* These indicators measure the distribution of overall importance or influence among nodes or edges within a network, considering attributes such as the number of direct connections (degree), the involvement of the node in the shortest paths (betweenness), or its association with other highly central nodes (eigenvector centrality). The mathematical expression for centralization (Z) is given by:

$$Z = \sum_i \frac{\max_j C_j - C_i}{(n-1) \cdot (n-2)} \tag{3.36}$$

  where Z denotes the centralization value of the network, C_i represents the centrality score of node i, $\max_j C_j$ is the maximum centrality score observed among all nodes, and n refers to the total number of nodes in the network.

Finally, for each prediction technique applied, the absolute difference between the assortativity values of the original and the predicted collaboration networks was also calculated to evaluate prediction accuracy at the structural level.

Chapter 4

Results

4.1 Results of the Ownership network analysis

4.1.1 Descriptive statistics

The ownership network of European companies was investigated within the timeframe of 2010 and 2018 as the yearly data was collected for all of these mentioned years. The Amadeus database was utilized, containing information on 23,381,325 companies. From this dataset, 1,872,272 companies were identified as parent companies or subsidiaries within the examined time frame. After data cleaning was performed, a final set of 1,620,340 distinct parent companies and subsidiaries was obtained. The investigated entities are associated with 1,435 NUTS 3 regions, which were used as nodes of the temporal network. The 87,708 ownership relationships identified between companies during the studied period are represented by the edges of the network. It should be noted that ownership relations occurring within the same NUTS 3 region resulted in self-loops.

The descriptive statistics for the main economic and technological indicators for the examined time period are presented in Table 4.1.

TABLE 4.1: Absolute indicators for ownership analysis

Description	2010	2011	2012	2013	2014	2015	2016	2017	2018	All
Mean value of P/L for period	1083	1122	1167	1262	1353	1473	1648	1874	2008	1.443
Mean value of P/L before taxes	1353	1411	1458	1493	1628	1757	1919	2193	2341	1.728
Mean value of cash flow	1426	1477	1568	1622	1725	1909	2045	2154	2231	1.795
Total number of employees	51791	56005	59190	63453	66819	72696	77410	80243	81375	67695
Total number of patents	26109	27088	27868	28275	28877	29065	24313	9753	1385	202733

The profit and loss (P / L) for the period statement (which means the net income, but in the Amadeus database the P / L for the period naming was used) shows the mean values in thousand € per year, calculated across all companies without aggregation at the NUTS 3 level. The mean value of P/L before taxes reflects the annual average for all companies, also in thousand €. In the cash flow line, the annual average cash flows for all companies are reported in thousand €. The number of employees represents the annual mean number of employees per company.

It can be observed that all indicators, except the patent data in the last three years, exhibit a generally increasing trend over time. The most probable reason for the difference in the Patstat-related data is explained in Section 3.1.4.

4.1.2 Null models for link prediction

Null models are designed to predict links. Simultaneously, through link prediction, a corresponding network structure is also generated. Figure 4.1 presents the fits of the null models, where **A** denotes the adjacency matrix of the original company ownership network CON, and **P** represents the adjacency matrix of the predicted networks. The figure shows on a log-log chart the comparison between the actual and predicted edges in the networks. The x-axis of the figures represents the original network-related links $A_{i,j}$, while on the y-axis the predicted links are presented $P_{i,j}$ as the matrices are flattened into vectors. The values represent the weight values for the edges from the original network on the x-axis and for the predicted ones on the y-axis, which are shown as blue crosses on the figures. The blue dots represent the same values for the self-loops. The red line is a visual aid that shows the perfect prediction. So, for example, when we see for a blue cross the value 1000 on the x-axis and 100 on the y-axis it means that the weight of the same edge is 1000 in the original network and 100 in the predicted one. The weights are calculated differently by the different models.

The model proposed by Newman and Girvan (2004) assumes a random network structure (Figure 4.1-a); however, it does not account for self-loops, and it fails to capture the distance-dependent probability of links between spatial nodes (i.e. NUTS 3 regions), as illustrated in Figure 4.2. In contrast, the model introduced by Expert et al. (2011) incorporates the non-linear dependency on the distance between nodes (Figure 4.1-b), thereby eliminating the appearance of clustered loops. The distance-dependent function $f(d)$ is compensated for through a spline function (see Eq. 3.4 and Figure 4.2). Nevertheless, the best fit, with an error of $\epsilon = 0.0080$, is achieved by the proposed GEN model.

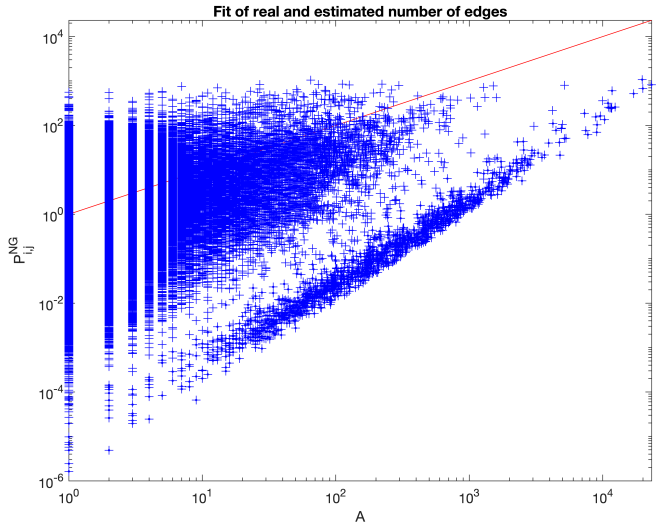
When applying Eq. (3.6), indicators with a variance inflation factor (VIF > 2.5) were excluded from the model. The adjusted R^2 value decreased only slightly as a result (see Table 4.2 and compare with Table B.1 in the Appendix B); thus, the assumptions of normality, homogeneity, and independence were considered satisfied for the refined model. As reference, the meaning of the used indicators table was copied here from Section 3.2.1 to Table 4.3.

Table 4.2 presents the results obtained from the proposed GEN model for the years under investigation. The table includes the estimated coefficients and their significance levels, along with the absolute errors of the fit, evaluated by comparing the original and predicted networks in terms of link prediction and centralities derived via the gravity model.

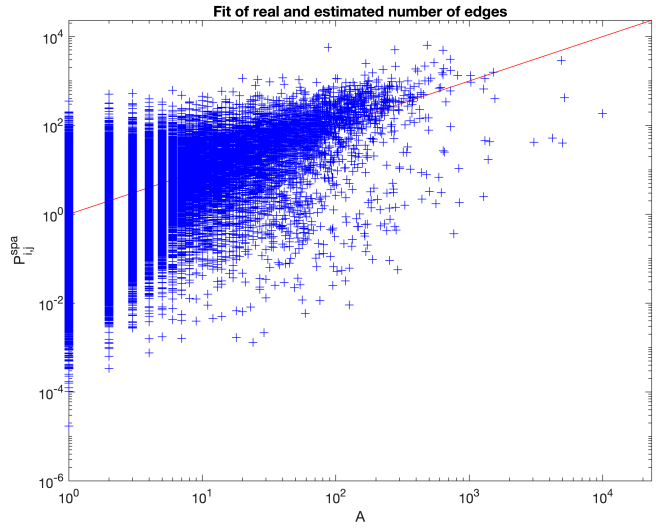
For instance, a coefficient of $\beta_{\mathrm{GDP}_{i,2015}} = -0.0390$ indicates that a 1% decrease in GDP in the source (i) region is associated with an expected increase of 0.0390% in the number of ownership links. As an example for understanding the results mentioned in the table, a positive significant coefficients on the source (subsidiary) side imply that an increase in the corresponding component may lead to an increase in ownership relations. Similarly, negative significant coefficients on the host side (parent companies) of the NUTS 3 regions suggest that an increase in these attributes may hinder investments and the establishment of new corporate sites.

Another good example is that $\beta_{\mathrm{FA}_{i,2018}} = 0.0111$ means on the parent company side that 1% increase in Fixed Assets within the NUTS 3 region indicates a 0.0111%

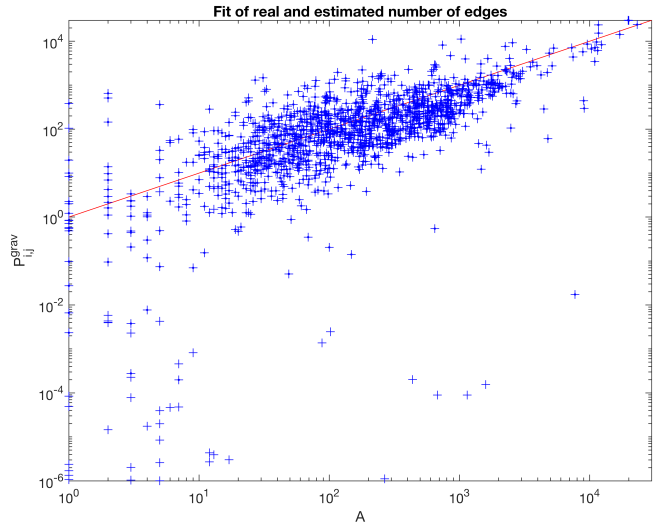
increase in willingness to establish a new subsidiary (somewhere). On the other hand, the $\beta_{\mathrm{FA}_{j,2018}} = -0.0199$ value on the subsidiary side, the 1% decrease of the Fixed Assets value in this region indicates a 0.0199% of increase the possibility of having a new subsidiary established in this region.



(a) Newman and Girvan (2004)'s model: $||\mathbf{A} - \mathbf{P}^{\mathrm{NG}}|| = \epsilon^{NG} = 0.0191$



(b) Expert et al. (2011)'s model: $||\mathbf{A} - \mathbf{P}^{\mathrm{spa}}|| = \epsilon^{spa} = 0.0112$



(c) GEN model: $||\mathbf{A} - \mathbf{P}^{\mathrm{grav}}|| = \epsilon^{grav} = 0.0080$

FIGURE 4.1: Fits of the different null models (2018)

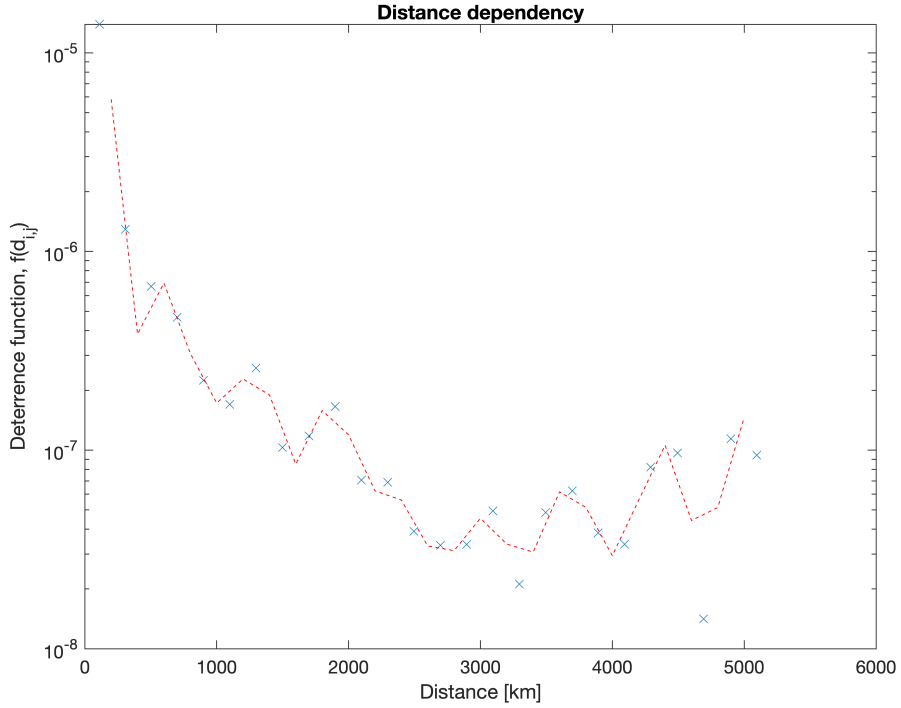


FIGURE 4.2: Distance deterrence function (2018)

TABLE 4.2: Gravity models results summary

(a) Regression coefficient values with significances

Coefficients	2010 β		2011 β		2012 β		2013 β		2014 β		2015 β		2016 β		2017 β		2018 β	
(Intercept)	1.5829	***	1.6187	***	1.5723	***	1.5766	***	1.6713	***	1.8964	***	1.6534	***	2.0972	***	2.1456	***
D_{ij}	-0.4726	***	-0.4725	***	-0.4727	***	-0.4729	***	-0.4735	***	-0.4717	***	-0.4718	***	-0.4711	***	-0.4725	***
SR_i	-0.1155	***	-0.1193	***	-0.0897	***	-0.1106	***	-0.1283	***	-0.1316	***	-0.1113	***	-0.1105	***	-0.0969	***
RB_i	-0.0717	***	-0.0551	***	-0.0695	***	-0.0603	***	-0.0616	***	-0.0670	***	-0.0910	***	-0.0747	***	-0.0866	***
RCB_i	0.0086		0.0079		-0.0035		-0.0062		-0.0214	***	0.0047		0.0001		-0.0020		0.0071	
FA_i	0.0206	***	0.0183	***	0.0159	***	0.0163	***	0.0128	***	0.0210	***	0.0177	***	0.0149	***	0.0111	***
CR_i	0.1236	***	0.1099	***	0.1219	***	0.1689	***	0.1986	***	0.1280	***	0.1765	***	0.1138	***	0.0795	***
CO_i	0.2193	***	0.2172	***	0.2215	***	0.2202	***	0.2209	***	0.2255	***	0.2287	***	0.2281	***	0.2299	***
GDP_i	-0.0023	*	-0.0019	*	-0.0016		-0.0013		-0.0023	*	-0.0390	***	-0.0258	***	-0.0285	***	-0.0223	**
PI_i	-0.0042	**	-0.0030	*	-0.0040	**	-0.0044	**	-0.0029	*	-0.0021		-0.0031	*	-0.0009		0.0026	
SR_j	-0.0054		-0.0099		0.0043		0.0017		0.0158		0.0594	***	0.0758	***	0.0260		0.0058	
RB_j	-0.0057		-0.0112		-0.0291	***	-0.0351	***	-0.0459	***	-0.0430	***	-0.0548	***	-0.0473	***	-0.0599	***
RCB_j	-0.0152	**	-0.0130	**	-0.0197	***	-0.0213	***	-0.0326	***	-0.0178	***	-0.0160	***	-0.0229	***	-0.0134	***
FA_j	-0.0122	***	-0.0120	***	-0.0114	***	-0.0111	***	-0.0123	***	-0.0086	***	-0.0107	***	-0.0159	***	-0.0199	***
CR_j	0.0464	**	0.0537	***	0.0514	***	0.0726	***	0.0811	***	0.0100		-0.0210		-0.0661	***	-0.0583	***
CO_j	0.2218	***	0.2205	***	0.2216	***	0.2223	***	0.2244	***	0.2304	***	0.2326	***	0.2327	***	0.2344	***
GDP_j	-0.0076	***	-0.0080	***	-0.0085	***	-0.0082	***	-0.0086	***	-0.0139	**	-0.0062		-0.0163	**	-0.0152	*
PI_j	-0.0105	***	-0.0096	***	-0.0091	***	-0.0101	***	-0.0095	***	-0.0139	***	-0.0139	***	-0.0091	***	-0.0089	***
Adj. R^2	0.4034	***	0.4029	***	0.4029	***	0.4032	***	0.4041	***	0.4026	***	0.4032	***	0.4019	***	0.4030	***
ϵ^{grav}	0.0078		0.0073		0.0078		0.0077		0.0076		0.0080		0.0081		0.0080		0.0082	

Values are significant at: * p=0.05. ** p=0.01. *** p=0.001 levels.

(b) Absolute errors of estimated centralities

Errors	2010	2011	2012	2013	2014	2015	2016	2017	2018
$\epsilon_{C_D^+}$	3.0414	5.7021	3.3355	3.5763	4.2312	3.8885	4.6273	6.5212	4.7929
$\epsilon_{C_D^-}$	4.5173	3.6933	5.1694	4.9229	4.9568	5.5349	4.1927	6.3334	5.5961
ϵ_{C_B}	142.7669	183.9388	139.3517	160.2476	161.9900	147.5215	167.2923	185.0560	182.4788
$\epsilon_{C_C^+}$	2.72E-06	3.25E-06	2.46E-06	2.22E-06	1.75E-06	2.77E-06	2.99E-06	1.91E-06	2.43E-06
$\epsilon_{C_C^-}$	4.62E-06	1.97E-06	4.71E-06	4.06E-06	3.60E-06	4.59E-06	2.77E-06	2.26E-06	3.77E-06
ϵ_{C_H}	1.98E-05	1.67E-05	2.06E-05	1.94E-05	1.85E-05	2.13E-05	1.81E-05	1.64E-05	2.06E-05
ϵ_{C_A}	1.42E-05	1.93E-05	1.28E-05	1.38E-05	1.39E-05	1.47E-05	1.45E-05	1.22E-05	1.53E-05
ϵ_{C_P}	2.83E-05	3.45E-05	2.62E-05	3.15E-05	3.18E-05	3.25E-05	2.23E-05	2.05E-05	3.00E-05

TABLE 4.3: Applied indicators for ownership network - reminder from Table 3.1

	v	Indicators	Description	Data source
Node dataset (NUTS 3 regional data)	m_1	TA	Total Assets	Amadeus
	m_2	SR	Solvency ratio (Asset based) (%)	Amadeus
	m_3	SH	Shareholders' funds	Amadeus
	m_4	RB	ROE using P/L before tax (%)	Amadeus
	m_5	RCB	ROCE using P/L before tax (%)	Amadeus
	m_6	PM	Profit margin (%)	Amadeus
	m_7	PLF	P/L for period	Amadeus
	m_8	PLB	P/L before tax	Amadeus
	m_9	OR	Operating revenue	Amadeus
	m_{10}	FA	Fixed Assets	Amadeus
	m_{11}	EN	Number of employees	Amadeus
	m_{12}	CR	Current ratio	Amadeus
	m_{13}	CF	Cash flow	Amadeus
	m_{14}	CO	Number of companies	Amadeus
	m_{15}	GDP	GDP/ capita in purchasing power priority	Eurostat
	m_{16}	PI	Patents	PATSTAT
Edges	i	FROM	The NUTS 3 ID of parent companies	Amadeus
	j	TO	The NUTS 3 ID of daughter companies	Amadeus
	$d_{i,j}$	Dist	Distance between regions	Eurostat
	$a_{i,j}$	OWN	Number of ownerships	Amadeus

Table 4.2(a) indicates that the applied model (see Eq. 3.6) is statistically significant, and the adjusted R^2 slightly exceeds 0.4. Among the independent variables examined, the coefficients for Fixed Assets (FA) are found to be strongly significant, with positive values on the source side (i) and negative values on the host side (j). This result suggests that parent companies generally possess higher levels of Fixed Assets compared to their subsidiaries.

The current ratio (CR), which measures liquidity as the ratio of current assets to current liabilities, exhibits high and significant coefficients on the source side. On the host side, these coefficients are smaller and remain positive only up to the year 2015. In contrast, the coefficients associated with the solvency ratio (SR) show an opposite trend in relation to the CR, indicating that parent companies are typically more liquid but less solvent than their subsidiaries.

The return on capital employed (ROCE), referred to as RCB in Table 4.2(a), serves as an indicator of operational efficiency. The coefficients for ROCE are significantly negative on both sides; however, this negative impact is more pronounced and statistically stronger for subsidiaries.

To assess the impact of economic and technological development at the regional level, the GDP per capita (GDP) and the annual number of patent applications (PI) were incorporated as additional variables. The coefficients associated with these indicators are negative for both the source and the host sides, although they are smaller in magnitude and less significant for the source regions. This pattern suggests that subsidiaries tend to be located in NUTS 3 regions characterized by lower GDP levels and fewer patent applications, relative to parent companies.

In addition, the number of companies (CO) within each NUTS 3 region was included as a control variable to account for the size of the regions. For this indicator,

the estimated coefficients were found to be highly positive on both sides, aligning with preliminary expectations. The coefficients corresponding to the distance between the parent and subsidiary companies are negative and remained relatively stable throughout the examined time period.

Finally, the original company ownership network was predicted using the applied model (see Eq. 3.6), and the mean absolute deviation between centrality measures calculated from the original and the predicted networks was evaluated. As shown in Table 4.2(b), these deviations are relatively consistent throughout the years under study.

4.1.3 Network property prediction

Since null models predict the links between nodes, the corresponding networks can also be predicted. Consequently, centralities can be calculated for the predicted networks as well. A good fit between the predicted and original centralities implies that the link predictions are accurate. Nevertheless, the discrepancies between the actual and predicted parameters offer additional insight into the structure of corporate networks.

Table 4.4 presents the mean absolute error of the centrality measures between the original and the predicted networks.

TABLE 4.4: Centralities prediction error

Prediction error of centralities	Random	Spatial	Gravity
In-degree, $\epsilon_{C_D^+}$	33.36681896	32.90683376	4.79292979
Out-degree, $\epsilon_{C_D^-}$	33.40637236	32.94787645	5.59613725
Betweenness,ϵ_{C_B}	170.04457052	169.95647946	182.47883742
In-closeness, $\epsilon_{C_C^+}$	0.00000940	0.00000919	0.00000243
Out-closeness, $\epsilon_{C_C^-}$	0.00000938	0.00000917	0.00000377
Hubs, ϵ_{C_H}	0.00002001	0.00002003	0.00002061
Authority, ϵ_{C_A}	0.00001703	0.00001704	0.00001532
PageRank, ϵ_{C_P}	0.00001782	0.00001781	0.00003000

Three types of networks can be predicted. The method proposed by Newman and Girvan (2004) generates a random network, where links are predicted using Eq. (3.3), and no organizing force is assumed. In this model, the edges between two regions are estimated proportionally to their numbers of incoming and outgoing connections. A spatial network is generated by the method of Expert et al. (2011), based on the model given in Eq. (3.4), which incorporates compensation for the distance dependency between nodes.

Although the prediction errors for these models are relatively low ($\epsilon^{NG} = 0.0191$ and $\epsilon^{spa} = 0.0112$), the absolute differences between the centralities remain quite similar. Notable improvements are achieved only with the gravity model. Due to the lower mean absolute error in link prediction ($\epsilon^{grav} = 0.0080$), the error associated with degree centrality is significantly reduced. However, for betweenness centrality and PageRank centrality, the prediction errors are observed to be larger.

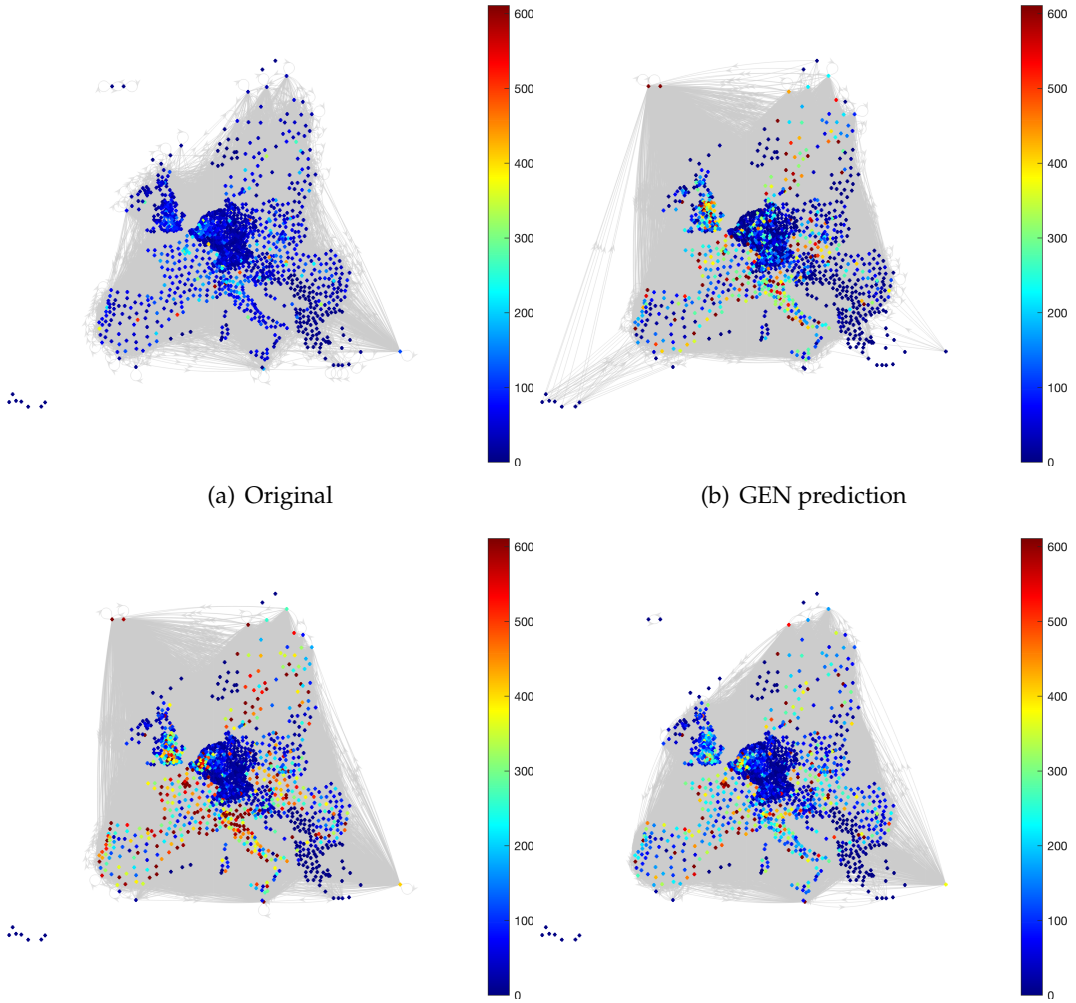
Table 4.5 illustrates an example that demonstrates the potential of modeling centralities. Better fits are achieved particularly in the case of in-degree and out-degree centralities. Table 4.5 lists the five regions that show the highest in-degree centralities, which correspond to the regions that are most attractive for the establishment of new subsidiary companies.

Table 4.5 demonstrates that the gravity-based GEN model predicts the ranks of the top five regions more accurately than the distance-dependent model. These regions are typically either capitals (e.g., Madrid, Rome, Warsaw) or major cities (e.g., Barcelona, Milan, Turin). The distance-dependent models, consistent with the higher prediction errors (ϵ), perform poorly in estimating these regions. This outcome suggests that, beyond accounting for geographical distances, economic, technological, and financial indicators must be incorporated into the null model to effectively predict the top role-player regions.

TABLE 4.5: Top 5 NUTS 3 regions in 2018 by in-degree centralities

C_D^-	Real network		GEN model			Spatial model		
Rank	NUTS 3	Name	NUTS 3	Name	Rank	NUTS 3	Name	Rank
1	ITC4C	Milan	ITC4C	Milan	(1)	'DK014'	Bornholm	(1291)
2	ITC11	Turin	PL911	Warsaw	(3)	'DK050'	Nordjylland	(589)
3	PL911	Warsaw	ES300	Madrid	(5)	'EE004'	Lääne-Eesti	(634)
4	ES511	Barcelona	ES511	Barcelona	(4)	'EE007'	Kirde-Eesti	(1081)
5	ES300	Madrid	ITI43	Roma	(7)	'EE008'	Lõuna-Eesti	(455)

Figure 4.3 displays the in-degree centralities of the NUTS 3 regions. In Figure 4.3(a), the results for the original network are presented, while Figures 4.3(b-d) depict the corresponding results for the predicted networks. To ensure a clear comparison across the different networks, the same color bar scale has been applied consistently to all visualizations.



(a) Original (b) GEN prediction

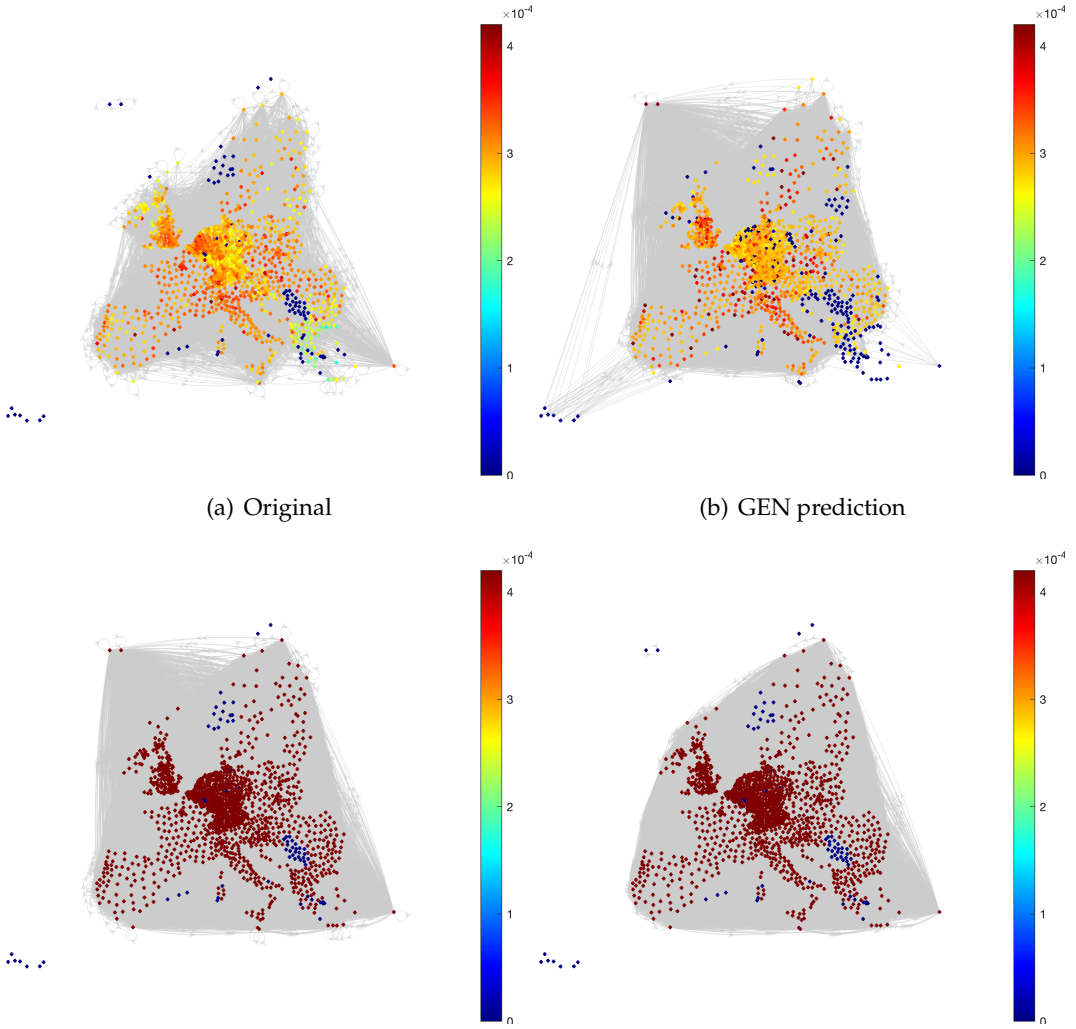
(c) Random network (d) Spatial network

FIGURE 4.3: The in-degree centralities for the predicted network structures (2018)

All network predictions reveal low in-degree centrality values for the United Kingdom, the Benelux countries (Luxembourg, the Netherlands, and Belgium), and Germany. Additionally, fewer nodes with high in-degree centrality are observed in the original network compared to the predicted networks. In-degree centralities are found to be overestimated, particularly in the random network and spatial network models, for both Southern and Central European countries. Consequently, the actual level of investment activity (as indicated by the establishment of corporate sites) in Southern and Central Europe is substantially lower than the levels predicted by any of the models. This finding suggests that investment in these regions is considerably less than what would be expected based on economic opportunities, including spatial, technological, and economic distances between regions.

Fig. 4.4 shows the closeness centralities of NUTS 3 regions. Fig. 4.4(a) shows the original network, and Figs. 4.3(b-d) show the predicted networks. The predicted networks use the same color bars as the original one to ensure an easy comparison of the results.

Fig. 4.4 shows that the in-closeness centralities are overestimated by both the random and spatial network models, and the best predictions are provided by the



(a) Original (b) GEN prediction

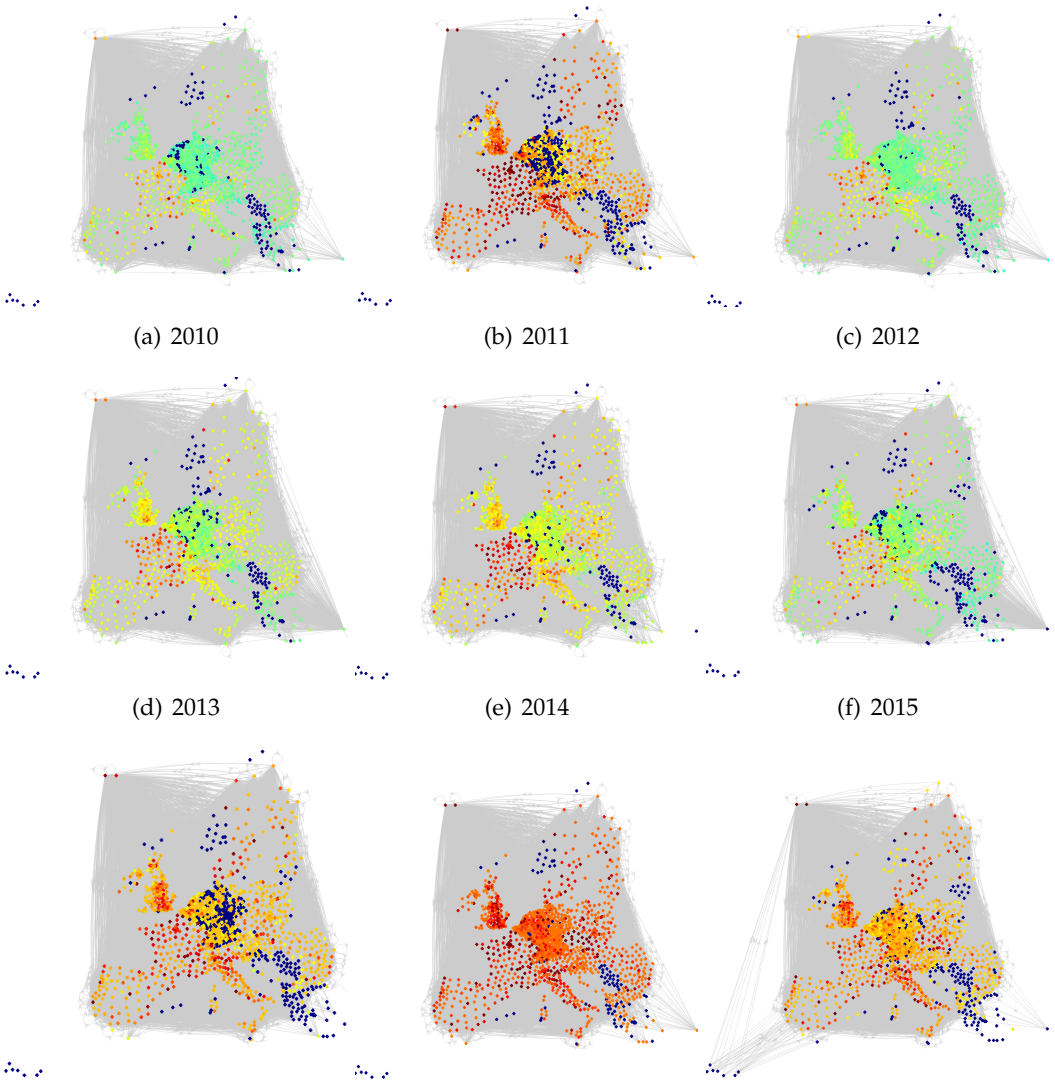
(c) Random network (d) Spatial network

FIGURE 4.4: In-closeness centralities for predicted networks (2018)

gravity model. It can be also identified by this Figure that for Serbian NUTS 3 regions, all of the models including the original one show low in-closeness centralities. Although for several southern England and eastern German areas predicted as playing important roles by the original and also by the gravity models.

Fig. 4.5 shows for all investigated years the in-closeness values of the multilayer network which was predicted by the GEN model. The in-closeness values reflect on the number of subsidiaries established in the NUTS 3 regions. On this Figure the yearly changes can be seen year-by-year as based on the in-closeness values, are increased in the UK and Germany for investments based on the ownership of the entities. The NUTS 3 regions can be identified that had more and more connections as targeted ones, which means more and more subsidiaries were established in these regions. From the Figure it can be seen that this kind of investments are not dominantly flowing out from the main countries of the European Union but remain inside. As the dominance of UK and Germany can be clearly identified in each year, these two main economies keep their capital mostly inside the country borders.

From this result it can be derived that even though, for example in Hungary, we can feel that several German companies establish great investments as subsidiary



(a) 2010 (b) 2011 (c) 2012

(d) 2013 (e) 2014 (f) 2015

(g) 2016 (h) 2017 (i) 2018

FIGURE 4.5: GEN predicted centralities between 2010-2018

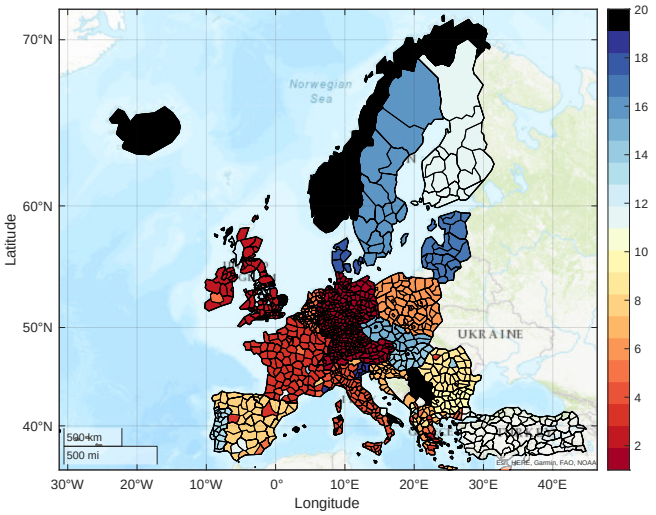
establishment within our country, even more of them could be possible. The reason why it is not happening can be a field of investigation about how this kind of investment could be supported for example by governmental assistance.

4.1.4 Identifying Economic communities

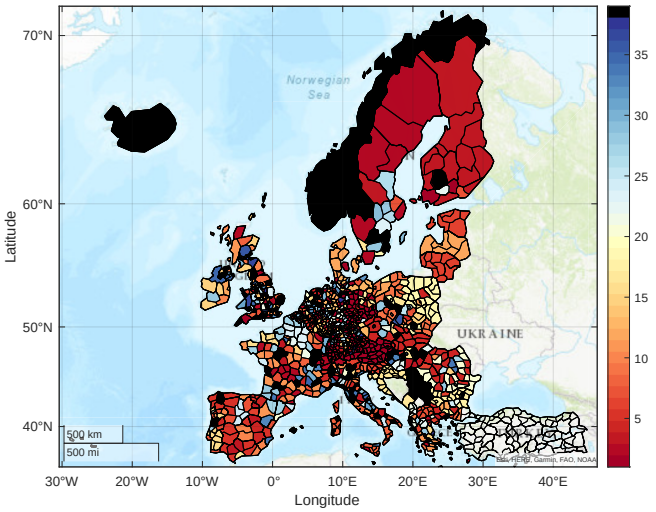
When modules are identified, groups of nodes that are more densely connected than predicted by the null models are determined. In the case of the model by Newman and Girvan (2004) (see Fig. 4.6-a), the modules correspond to sets of NUTS 3 regions that exhibit stronger internal connectivity compared to their inter-module connections.

The spatial model (see Fig. 4.6-b) and the proposed GEN model (see Fig. 4.6-c) already incorporate spatial and economic characteristics during the prediction process. As a result, distance-dependent modules reveal communities in which the internal regional connections are stronger than those expected by the distance-compensated model. When modules are determined based on predictions from the GEN model, the resulting economic communities indicate sets of regions with connection strengths that exceed those explained by spatial, economic, or technological attributes alone.

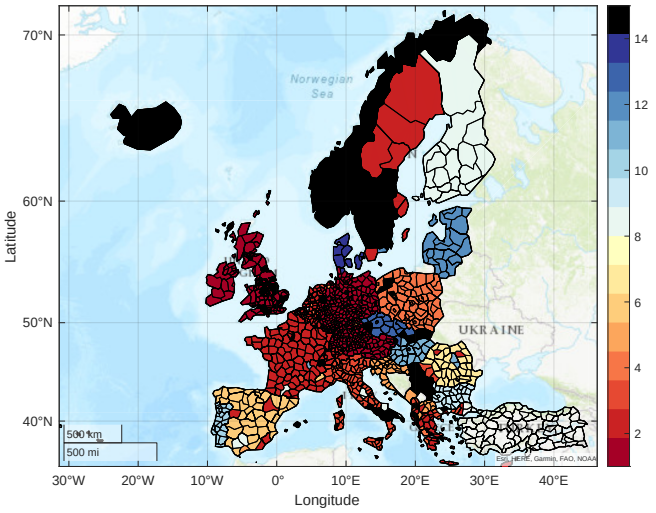
Figure 4.6 presents the modules detected according to the different null models. Modules represented by lower values (in reddish tones) include a larger number of regions, whereas those in bluish shades consist of fewer regions and correspond to higher numerical labels. The modules shown in black contain only one region each. Due to the absence of data for Turkey in the data set used, the Turkish NUTS 3 regions are displayed in white.



(a) Newman and Girvan (2004)'s modules

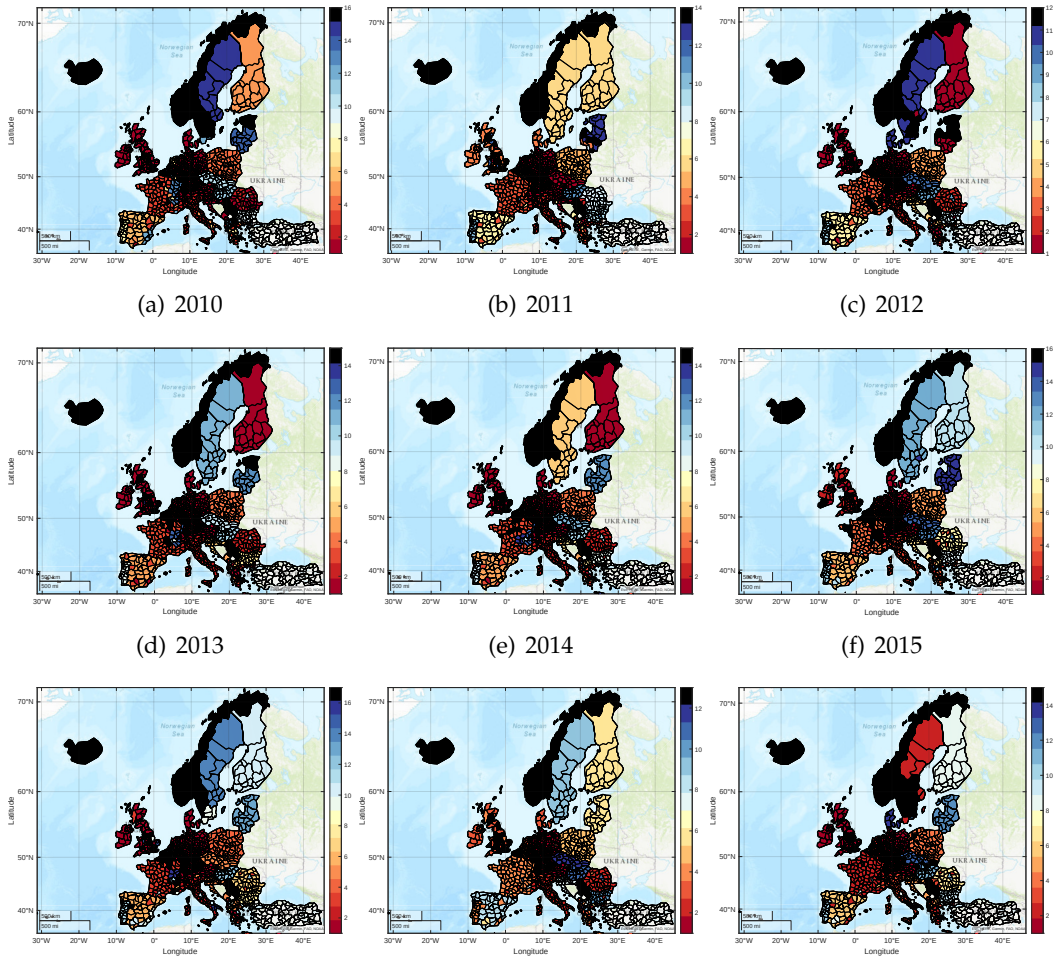


(b) Distance-dependent modules



(c) Economic communities (2018) - GEN-model

FIGURE 4.6: Modules of NUTS 3 regions.



(a) 2010 (b) 2011 (c) 2012

(d) 2013 (e) 2014 (f) 2015

(g) 2016 (h) 2017 (i) 2018

FIGURE 4.7: Layers as years (2010-2018) of the found economic modules

As shown in Gadár et al. (2018), gravity-based modules are capable of delineating investment catchment areas that can transcend administrative boundaries. However, the observation that the gravity model predominantly reproduces administrative divisions implies that the formation of parent-subsidiary corporate relationships remains largely constrained by these boundaries. This indicates that economic communities are primarily structured within national borders. Moreover, it is evident that most regions in Great Britain and Germany, as well as France and Northern Italy, form distinct and internally cohesive economic blocks.

Figure 4.7 illustrates the identified economic modules within the multilayer network, where each layer corresponds to a specific year. This Figure shows the benefit of using the multilayer approach as each layer shows a particular year. There are no connections between the different layers, so all years are represented based on the data related to the particular time frame. When identifying the communities within each layer separately, the Figure can be drawn. Based on the fact that all the coloring is used in the same way in each layer as it was used in Fig. 4.6, the temporal change can be visually identified. Instead, all layers show a very similar formation and coloring, which means that no significant change between the layers can be identified, and consequently we can say that the modules based on the GEN model are stable

in time within the investigated time frame between 2010 and 2018.

4.2 Results of the H2020 Collaboration network analysis

4.2.1 Descriptive statistics

Table 4.6 provides an overview of the main structural metrics that describe the collaboration network established under the Horizon 2020 framework program.

TABLE 4.6: Descriptive statistics of network properties

Property	Values
Vertices	20,172
Edges	237,084
Number of components	4,363
Average distance (vertices)	3.37862
Average distance (km)	977.65710
Density	0.00045
γ	2.56377
Assortativity	-0.05226
Degree of centralization	0.12146
Betweenness centralization	0.10415
Eigenvector centralization	0.99259
Modularity	0.61364

The network is composed of 20,172 vertices (nodes), each representing an entity - specifically organizations of at least a small size - that have participated in Horizon 2020 R&D&I projects and are identified by a European Bvd ID Number in the Orbis database. There are 237,084 edges in the network, and each edge denotes a collaborative link formed through joint project participation.

A significant feature of the network is the presence of 4,363 distinct components. Combined with the low density value of 0.00045, this points to a structure that is far from fully interconnected, instead consisting of many smaller, isolated sub-networks. Despite this fragmentation, the average shortest path length between nodes within components is relatively low (3.37862), indicating that entities are closely linked within their respective sub-networks. In contrast, when considering the physical (geographical) separation between entities, the average distance increases to 977.65710 km. This suggests that collaborations frequently occur across large geographic spans, reflecting the international nature of the network.

The network's degree distribution, reflected by a γ value of 2.56377, aligns with a power-law distribution. This characteristic implies that a small subset of organizations serve as highly connected hubs, while most entities maintain only a limited number of connections. Such a pattern is typical of scale-free networks Meng and Zhou (2023) and Voitalov et al. (2019), where connectivity is highly uneven.

The value of γ comes from the following so-called power-law distribution formula Barabási (2016): $P_k \sim k^{-\gamma}$, where P_k is the probability that a node has a degree k, k is the degree of a node and γ is the *power-law exponent*.

The negative assortativity coefficient (-0.05226) indicates a slight tendency for highly connected organizations to collaborate with those that have fewer connections, rather than forming ties primarily with other well-connected entities. However, since this negative value is modest, the inclination for organizations to work with dissimilar partners is present, but weak; the network does not exhibit strong assortative or disassortative mixing.

Further insights can be drawn from the centralization measures. The degree centralization (0.12146) is rather low, suggesting that the network does not revolve around a few central organizations. Similarly, the betweenness centralization (0.10415) indicates that no single entity has a dominant role in mediating collaborations or controlling the flow of information. On the other hand, the eigenvector centralization is notably high (0.99259), pointing to the existence of certain organizations that are particularly influential due to their connections with other central entities.

Finally, the modularity value of 0.61364 reveals a pronounced modular structure within the network. This high modularity signifies the existence of several well-defined communities or clusters, where organizations are more densely connected within their own group than with those outside it.

Table 4.7 displays key organizational metrics stratified by firm size categories.

TABLE 4.7: Key Organizational Metrics Across Firm Size Categories

Metric\Firm Size	**Small**	**Medium**	**Large**	**Very Large**
Total Assets (TA)	7,180	31,834	38,512	3,009,231
Operating Revenue (OR)	4,782	29,133	11,258	1,789,507
Profit/Loss Before Tax (PLB)	399	650	1,038	126,959
Multi-project Membership (MULTI)	0.0707	0.0872	0.0666	0.0961
Program Membership (PROG)	0.4390	0.4450	0.4210	0.4410
Number of Firms (N)	6,647	3,528	7,038	2,959

Table 4.7 reveals marked disparities in resource allocation profiles across firm sizes. Total Assets (TA), operating revenue (OR), and profit/loss before tax (PLF) exhibit significant heterogeneity, reflecting divergent operational scales. For example, very large firms report TA values three orders of magnitude greater than small firms (3,009,231 vs. 7,180). In contrast, the multi-project (MULTI) and program (PROG) membership metrics show minimal variation, with all size categories clustering close to 0.07–0.10 and 0.42–0.45, respectively.

Stratified analysis indicates that very large and medium companies exhibit slightly higher multi-project engagement (MULTI = 0.0961 and 0.0872), although these values remain consistently low across all categories. Program membership (PROG) demonstrates greater uniformity, with all groups maintaining values between 0.42 and 0.45. This consistency, coupled with the high values N (particularly for small and large firms), suggests a systematic preference for collaborative engagements with partners who have prior project experience, regardless of organizational size.

4.2.2 Findings from Generic Machine Learning Approaches

TABLE 4.8: Group mean in LDA, the exponent of coefficients (Exp(β)), and significances in LogR. (Variable groups: c: Coorporate; C: Collaboration; E: Economy; T: Technology)

		Group mean (LDA)		LogR		**Group mean** (LDA)		LogR
	Partners	0	1	Exp(β)	**Coordinators**	0	1	Exp(β)
Coorporate (c)	**FROM_TA**	−0.0943	0.1243	1.0000 ∗ ∗	**TO_TA**	−0.0980	0.1291	1.0000 ∗ ∗
	FROM_SR	0.0357	−0.0470	0.9986 ∗ ∗∗	**TO_SR**	0.0354	−0.0467	0.9986 ∗ ∗∗
	FROM_SH	−0.0946	0.1246	1.0000 ∗ ∗	**TO_SH**	−0.0997	0.1313	1.0000 ∗ ∗
	FROM_RB	−0.0309	0.0407	1.0005 ∗ ∗∗	**TO_RB**	−0.0313	0.0412	1.0005 ∗ ∗∗
	FROM_RCB	−0.0284	0.0375	1.0004	**TO_RCB**	−0.0297	0.0392	1.0004
	FROM_PM	−0.0382	0.0503	1.0020 ∗ ∗∗	**TO_PM**	−0.0407	0.0537	1.0021 ∗ ∗∗
	FROM_PLF	−0.0851	0.1122	1.0000	**TO_PLF**	−0.0881	0.1160	1.0000
	FROM_PLB	−0.0852	0.1123	1.0000	**TO_PLB**	−0.0875	0.1153	1.0000
	FROM_OR	−0.1036	0.1365	1.0000∗	**TO_OR**	−0.1075	0.1417	1.0000
	FROM_FA	−0.0870	0.1146	1.0000 ∗ ∗∗	**TO_FA**	−0.0908	0.1197	1.0000 ∗ ∗
	FROM_EN	−0.1079	0.1421	1.0000	**TO_EN**	−0.1144	0.1507	1.0000
	FROM_CR	0.0177	−0.0233	0.9996	**TO_CR**	0.0170	−0.0224	0.9981
	FROM_CF	−0.0837	0.1103	1.0000	**TO_CF**	−0.0880	0.1160	1.0000
	FROM_Size	−0.1702	0.2242	1.1938 ∗ ∗∗	**TO_Size**	−0.1831	0.2413	1.2054 ∗ ∗∗
(E)	**FROM_GDP**	−0.0863	0.1137	1.0000	**TO_GDP**	−0.0937	0.1234	1.0000
	FROM_CI	−0.0352	0.0464	0.9941 ∗ ∗∗	**TO_CI**	−0.0368	0.0485	0.9938 ∗ ∗∗
	FROM_Pov	0.0195	−0.0257	1.0008	**TO_Pov**	0.0185	−0.0244	1.0013
(T)	**FROM_PI**	−0.1063	0.1400	1.0000	**TO_PI**	−0.1127	0.1485	1.0000
(C)	**FROM_EC**	−0.1725	0.2273	1.0000 ∗ ∗∗	**TO_EC**	−0.1831	0.2412	1.0000 ∗ ∗∗
	FROM_MULTI	−0.1857	0.2446	17.0223 ∗ ∗∗	**TO_MULTI**	−0.1945	0.2562	20.1564 ∗ ∗∗
	FROM_PROG	−0.0845	0.1113	0.5638 ∗ ∗∗	**TO_PROG**	−0.0880	0.1160	0.5737 ∗ ∗∗
(C)	**Edges**	0	1	Exp(β)				
	Dist	0.1142	−0.1504	0.9998 ∗ ∗∗				
	FP7	−0.0413	0.0544	72.2008				

Table 4.8 summarizes the standardized group means for linked (1) and unlinked (0) organizational nodes, along with the discriminant function coefficients stratified by partnership and coordination roles. The table differentiates between vertex-level indicators (for example, partner / coordinator status) and edge-level metrics (for example, geographical distance Dist and FP7 collaboration history). The variables are categorized into four thematic groups: corporate (c), economic (E), technological (T), and collaboration (C) variables.

The analysis of corporate variables reveals that organizations engaging in collaborative activities generally exhibit more favorable characteristics. These include higher values of Total Assets (TA) (as evidenced by the group means for both FROM_TA and TO_TA), Fixed Assets (FAs), improved Shareholder Funds (SH), ROE using P/L before tax (%) (RB), and ROCE using P/L before tax (%) (RCB), as well as increased Profit margin (%) (PM), Operation revenue (OR), P/L before tax (PLF), and P/L for period (PLF). Notably, only two corporate indicators, Current ratio (CR) and Solvency Ratio (asset based) (%) (SR), are observed to be lower among collaborating firms. Furthermore, when focusing on coordinators, the disparity in group means is even more pronounced, underscoring the heightened importance of robust corporate characteristics for organizations assuming coordinator roles.

Favorable conditions extend beyond internal resources to encompass the broader economic environment. Firms involved in collaborations tend to be located in regions with stronger economic indicators, such as higher GDP, reduced poverty rates,

and a more sustainable economy (reflected in a higher CI). The elevated patent values observed in collaborative organizations further highlight the significance of technological capability, indicating that technological background is a crucial factor in fostering collaboration.

From a network perspective, the propensity to initiate collaborations is linked to structural network features. As discussed previously, the network demonstrates scale-free properties, with project contributions and funding predominantly concentrated among a limited number of entities. This is reflected in the tendency of organizations to seek partnerships with those that exhibit higher Earned contribution (EC). Entities with extensive experience in managing multiple projects (MULTI) and a history of successful project participation (PROG) are more likely to engage in future collaborations.

Edge-level collaboration metrics indicate that the likelihood of cooperation decreases as the geographical distance between partners increases. In contrast, a history of previous collaboration under previous framework programs significantly enhances the probability of renewed cooperation. Upon dividing the dataset into training and testing subsets, the model achieved an accuracy of 0.77 and an F1 score of 0.85 in the test set, which is considered a strong performance for a general-purpose algorithm. The application of the QDA method yielded even higher accuracy (0.82) and an F1 score of 0.89, suggesting the presence of non-linear relationships between variables. Nevertheless, LDA still delivers satisfactory accuracy and offers valuable insights into the characteristics of the independent variables.

Table 4.8 further details the statistical significance of the explanatory variables. For both partners (FROM) and coordinators (TO), the corporate indicators RCB, PLF, PLF, Number of employees (EN), CR, and Cash flow (CF) were statistically insignificant. Among economic indicators, only the corruption index demonstrated significance, while all collaboration metrics except the previous collaboration history were significant. Due to the disparate scaling between explanatory and response variables, the odds ratios approximated 1. Consequently, the group mean values derived from LDA proved to be more interpretable than the coefficients, particularly since the minor values of β caused sign reversals relative to the standard LDA outputs. However, three significant variables yielded noteworthy interpretations.

Consistent with the findings of LDA, larger organizations exhibited a higher propensity for collaboration. The LogR model quantified this relationship, revealing that larger firms had a 19.38% increased likelihood of acting as partners and a 20.54% greater probability of assuming coordinator roles. The parallel project engagement metric (MULTI) produced the strongest odds ratio: organizations involved in concurrent collaborations were 17 times more likely to become partners and 20 times more likely to serve as coordinators in new projects.

In particular, the coefficient for program participation history (PROG) was negative, suggesting that previous success in similar projects does not systematically predict future collaboration opportunities. This finding appears to contradict the results of LDA, which indicated that partners frequently participate in projects based on previous initiatives. Although the coefficient for previous collaboration under previous framework programs was numerically large, its statistical insignificance (attributable to low prevalence in the dataset) limits interpretability.

The LogR method outperformed other techniques, achieving an accuracy of 0.83 and an F1 score of 0.90. However, like LDA and QDA, it remains constrained by an inability to address multicollinearity among predictors. This limitation motivated the supplementary use of non-generic, black-box machine learning approaches.

4.2.3 Analysis of Feature Importance Using Random Forest-Based Approaches

To assess the importance of variables in link prediction tasks, RF algorithms were utilized. The variable selection was performed using the Boruta algorithm. Following this selection process, both standard and parameter-optimized (fine-tuned) versions of the RF method were applied.

Figure 4.8 presents the results of the Boruta-based feature selection procedure.

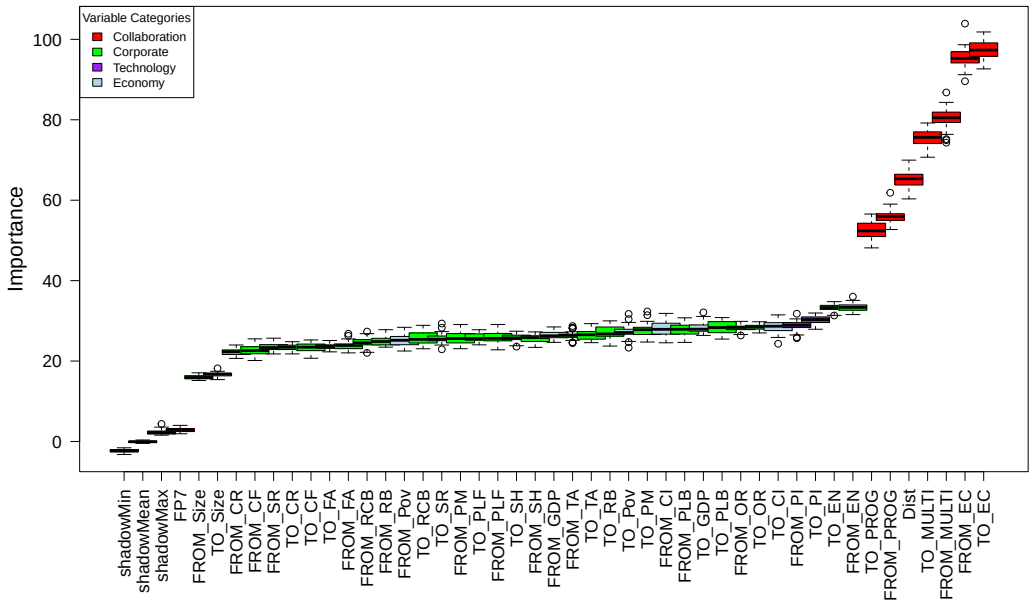
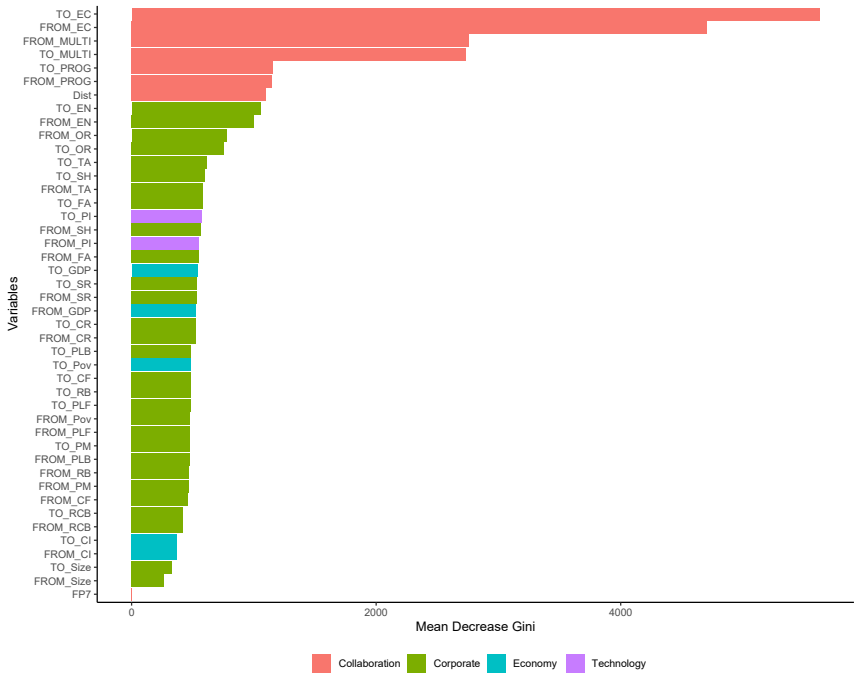


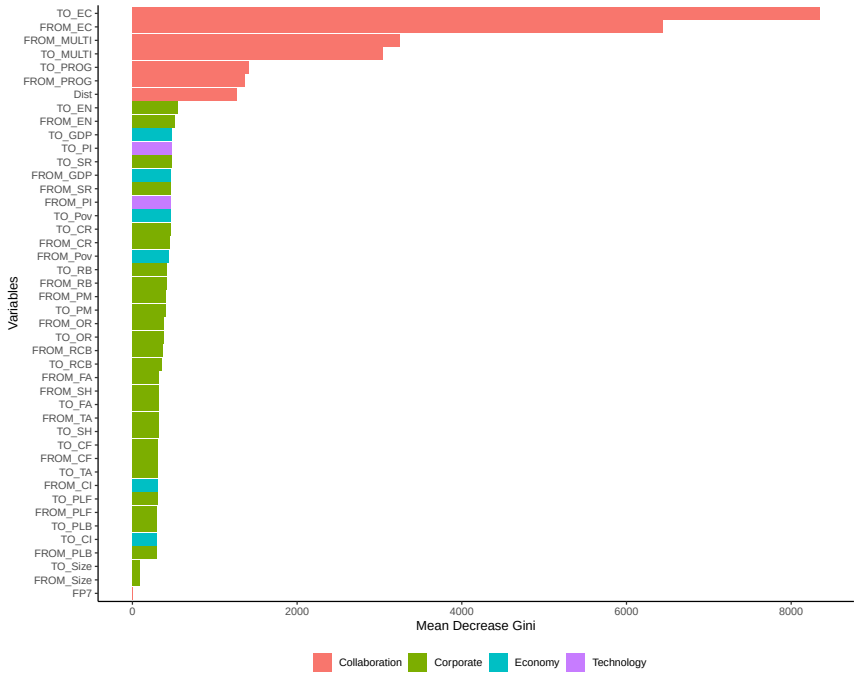
FIGURE 4.8: Feature importance as determined by the Boruta algorithm

Unlike the results obtained from LogR, the Boruta algorithm identified all variables as important, as their importance scores significantly exceeded those of the randomly generated (shadow) features. Among these, the most influential predictors were the earned contribution (EC) and the average rate of parallel project completion (MULTI). This outcome is consistent with the principle of preferential attachment or cumulative advantage, indicating that organizations that already have extensive connections and a history of multiple simultaneous projects are more likely to attract additional collaborations. In contrast, entities with fewer existing connections face greater challenges in establishing new collaborations, a barrier that appears to be more substantial than even geographical distance, which ranked as the fifth most significant factor.

Collaboration-related variables, specifically PROG, were ranked as the sixth and seventh most important features, underscoring the relevance of previous successful project experience in fostering new partnerships. Although the analysis LogR suggested that previous experience does not necessarily facilitate the continuation of specific research lines, the results LDA indicated a higher participation rate in such programs among collaborating organizations.



(a) No parameter tuning (Accuracy = 0.86, F1 score = 0.91)



(b) After parameter tuning (Accuracy = 0.86, F1 score = 0.91)

FIGURE 4.9: Feature importance values from RF models

Variables reflecting corporate, technological, or economic attributes only appeared from the eighth position onward. Among corporate indicators, the number of employees (EN) emerged as the most significant, highlighting the importance of organizational resources. The number of patents (PI) served as a proxy for technological capacity, while GDP per capita represented the economic context. The final relevant variable was the prior collaboration between organizations in the previous

framework program (7th Framework Programme (FP7)). In particular, the Boruta algorithm did not classify any feature as irrelevant.

Figure 4.9 illustrates the feature importance scores derived from the algorithm RF, both prior to parameter optimization (Fig. 4.9(a)) and following parameter tuning (Fig. 4.9(b)). Distinct categories of indicators are visually differentiated by color. Feature importance is quantified by the mean decrease in the Gini index, which reflects the average reduction in Gini impurity contributed by a given variable when it is utilized to split a decision node within the ensemble of trees.

Consistent with the Boruta results (see Fig. 4.8), Figure 4.9 demonstrates that the top seven predictors of link formation are collaboration-related indicators. Beyond earned contribution (EC), indicators that capture the completion structure of the project, such as MULTI and PROG, emerge as the most influential in the creation of new collaborative ties, surpassing even geographical distance in importance. Although parameter tuning of the RF model resulted in negligible improvements in accuracy and F1 score, it did alter the relative importance of some variables, bringing the results into closer alignment with those obtained using the Boruta method (compare Fig. 4.8 and Fig. 4.9(b)).

After tuning, the most prominent corporate predictors were the number of employees (EN), GDP per capita, and the number of patents (PI). In particular, the number of employees (EN) was of the highest importance among the corporate variables, while the organizational size was comparatively less influential within this group. These findings underscore that R&D&I collaborations are typically dependent on human capital, implying that organizations with ample human resources are better positioned to participate in collaborative projects.

4.2.4 Black-box prediction methods results

Although the RF approach is classified as a black-box method, it still allows the interpretation of variable importance. Conversely, when the primary objective is to maximize the accuracy of edge prediction and to employ robust techniques that remain effective under less restrictive conditions, the application of non-generic black-box methods is warranted. The performance metrics of both generic (LDA, QDA, and LogR) and non-generic algorithms, including RF, SVM, and XGBoost, in the test dataset, are presented in Table 4.9.

TABLE 4.9: Results of link prediction on the test dataset. (T: Tuned method)

	LDA	QDA	LogR	SVM	SVM (T)	RF	RF (T)	XGBoost	XGBoost (T)
Accuracy	0.771900	0.817738	0.832748	0.844329	0.820240	0.856434	0.859057	0.853206	**0.863374**
Kappa	0.379527	0.371745	0.472292	0.535281	0.498282	0.587077	0.610799	0.580088	**0.620298**
Accuracy (Lower)	0.766624	0.812874	0.828043	0.839755	0.815401	0.852006	0.854662	0.848738	**0.859035**
Accuracy (Upper)	0.777112	0.822527	0.837376	0.848822	0.825003	0.860778	0.863367	0.857590	**0.867627**
Accuracy (Null)	0.767704	0.767704	0.767704	0.767704	0.767704	0.767704	0.767704	0.767704	0.767704
Accuracy PValue	0.059524	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
McNemar PValue	0.000000	0.000000	0.000000	0.000000	0.267566	0.000000	0.000020	0.000000	0.012707
Sensitivity	0.838116	**0.962262**	0.939399	0.924997	0.880952	0.917429	0.901556	0.913014	0.907180
Specificity	0.553066	0.340108	0.480285	0.577731	0.619594	0.654855	**0.718603**	0.655550	**0.718603**
Pos Pred Value	0.861062	0.828154	0.856602	0.878632	0.884439	0.897799	0.913706	0.897541	**0.914195**
Neg Pred Value	0.508301	**0.731689**	0.705717	0.699769	0.611626	0.705860	0.688353	0.695156	0.700830
Precision	0.861062	0.828154	0.856602	0.878632	0.884439	0.897799	0.913706	0.897541	**0.914195**
Recall	0.838116	**0.962262**	0.939399	0.924997	0.880952	0.917429	0.901556	0.913014	0.907180
F1	0.849434	0.890185	0.896092	0.901219	0.882692	0.907508	0.907590	0.905211	**0.910674**
Prevalence	0.767704	0.767704	0.767704	0.767704	0.767704	0.767704	0.767704	0.767704	0.767704
Detection Rate	0.643425	**0.738732**	0.721180	0.710124	0.676310	0.704313	0.692128	0.700924	0.696445
Detection Prevalence	0.747246	0.892023	0.841908	0.808215	0.764677	0.784489	0.757495	0.780939	0.761813
Balanced Accuracy	0.695591	0.651185	0.709842	0.751364	0.750273	0.786142	0.810080	0.784282	**0.812892**

Table 4.9 shows that the highest predictive accuracy (0.863374) is achieved by the tuned XGBoost method, closely followed by the optimized RF model (accuracy = 0.859057). These methods also exhibit the highest kappa values, reflecting strong agreement between predicted and observed classifications. In terms of sensitivity (true positive identification), the QDA and tuned SVM methods perform optimally, with scores of 0.962262 and 0.924997, respectively. However, these approaches are associated with reduced specificity (true negative identification).

The tuned RF and tuned XGBoost models yield the highest positive predictive value (proportion of true positives among positive predictions) alongside the superior precision, recall and F1 scores. Consequently, these methods are identified as the most effective for link prediction in the analyzed collaboration network, achieving maximum accuracy, balanced accuracy, and robust performance in complementary metrics.

All evaluated methods significantly outperform the null accuracy baseline (see Section 3.3.2 - 0.767704), which represents the performance of a classifier that exclusively predicts the majority class. This confirms the non-trivial predictive capacity of the algorithms tested. The comparatively lower performance of generic methods such as LDA and non-generic approaches such as SVM suggests the presence of nonlinear relationships between explanatory and response variables.

The SVM algorithm was tested using multiple kernel functions, with the radial kernel and the C-classification approach delivering the highest accuracy. However, parameter optimization provided only marginal improvements across all methods, indicating limited gains from hyperparameter tuning in this context.

4.2.5 Comparative Analysis of Network Structure Prediction Performance

The predictive performance of the link estimation models was evaluated across the entire collaboration network. The structural discrepancies between the original (G) and predicted ($\widehat{G}$) networks are quantified in Table 4.10.

TABLE 4.10: Prediction accuracy and structural discrepancies between original (G) and predicted ($\widehat{G}$) collaboration networks across models

	Network	LDA	QDA	LogR	SVM	SVM (T)	RF	RF (T)	XGBoost	XGBoost (T)
	Accuracy (all)	0.718438	0.710722	0.775996	0.834299	0.875079	0.965371	0.965989	0.965583	0.928062
Structural discrepancies	Nodes	0	0	0	0	0	0	0	0	0
	Edges	-15208	-25859	-15199	-7349	-2514	-634	**-37**	-257	65
	Number of components	72	80	61	59	39	6	5	**4**	12
	Assortativity	-0.011348	-0.012460	-0.010368	-0.007606	-0.006570	-0.001907	**-0.001688**	-0.001850	-0.003846
	Density	-0.000037	-0.000064	-0.000037	-0.000018	-0.000006	-0.000002	**0.000000**	-0.000001	**0.000000**
	Average path length	-0.133734	-0.076950	-0.090980	-0.118293	-0.111480	**-0.023049**	-0.026438	-0.027690	-0.068630
	Degree centralization	0.016423	0.005245	0.010920	0.014792	0.015103	**0.003745**	0.004784	0.004835	0.011725
	Betweenness centralization	0.008601	-0.000696	0.002107	0.008538	0.008352	**0.001353**	0.002279	0.002573	0.007204
	Eigenvector centralization	0.001202	0.001095	0.000603	0.000842	0.000850	**0.000205**	0.000234	0.000231	0.000568
	Modularity	-0.009708	-0.012783	-0.012964	-0.002644	-0.006250	-0.002112	0.002740	-0.001129	**-0.000958**
Avg abs. diff.	Degree centrality	2.338588	2.835713	2.069601	1.485227	1.074658	0.320742	**0.304680**	0.325104	0.648721
	Normalized betweenness centrality	0.202733	0.145354	0.140855	0.170489	0.158315	**0.038794**	0.043452	0.046068	0.110242
	Eigenvector centrality	0.001912	0.001736	0.001493	0.001430	0.001455	**0.000483**	0.000510	0.000520	0.001088

The QDA method exhibited the lowest overall accuracy (0.710722), and all generic algorithms achieved sub-0.78 accuracy scores. All methods overestimated edge counts, as evidenced by negative discrepancies between G and $\widehat{G}$. Consequently, predicted networks were characterized by fewer disconnected components and marginally higher density.

The predicted networks exhibited structural parameters closely aligned with those of the original network, although a systematic reduction in disassortative mixing and network concentration was observed (negative assortativity differences, positive centrality differences). The mean absolute discrepancy in the values of DC suggests that the top performing methods (RF, XGBoost) estimate organizational partnership counts within ±1 partner. The most pronounced discrepancy was observed in normalized BC estimates, with LDA yielding the largest deviation (0.202733). Eigenvector centrality distributions were accurately replicated across methods, with maximal average absolute differences below 0.2%.

Parameter tuning in non-generic methods provided negligible improvements in prediction fidelity, with certain structural metrics, such as centralization measures, showing marginal degradation.

Given that link prediction accuracies were found to be satisfactory not only on the test and training datasets, but also across the entire network, and considering that the discrepancies in structural parameters were minimal, the prediction of collaborative ties and the identification of organizations occupying central roles in the collaboration network can be regarded as feasible. The optimal non-generic method achieved an overall accuracy of 0.928, indicating that 92.8% of link predictions were correct. The results for the most collaborative organizations, as determined by predicted degree centrality, are presented in Table 4.11.

TABLE 4.11: Top collaborators by their degree of centrality (links) and predictions based on generic (LDA, QDA, LogR) and non-generic methods. (T: Tuned method)

Ord.	Org. ID	Org. Name	Orig. DC	LDA	QDA	LogR	SVM	SVM (T)	RF	RF (T)	XGBoost	XGBoost (T)
1	DE8170003303	FRAUNHOFER-GESELLSCHAFT ZUR FOERDERUNG DER ANGEWANDTEN FORSCHUNG EINGE-TRAGENER VEREIN	4918	4257	4709	4479	4322	4309	**4767**	4725	4723	4445
2	BE0419052173	INTERUNIVERSITAIR MICRO-ELECTRONICA CENTRUM	1870	1471	1736	1567	1499	1474	**1761**	1744	1739	1565
3	DK30060946	DANMARKS TEKNISKE UNIVERSITET	1687	1464	1602	1552	1497	1470	**1627**	1620	1620	1532
4	BE0425260668	KATHOLIEKE UNIVER-SITEIT TE LEUVEN	1455	1449	1454	1451	1451	1450	**1455**	1454	1454	1453
5	DE5030021537	DEUTSCHES ZEN-TRUM FUER LUFT- UND RAUMFAHRT E.V.	1372	1198	1316	1258	1216	1203	**1335**	1326	1318	1261
6	DK31119103	AARHUS UNIVERSITET	1233	1137	1172	1130	1127	1114	1186	1180	**1187**	1157
7	NL09098104	STICHTING WAGENIN-GEN RESEARCH	1200	1103	1167	1115	1107	1107	**1185**	1179	1182	1134
8	NO919303808	SINTEF AS	1191	1094	1158	1108	1097	1087	**1167**	1162	1162	1124
9	BE0248015142	UNIVERSITEIT GENT	1150	1064	1146	1104	1087	1083	**1123**	1115	1114	1110
10	AT9110045499	AIT AUSTRIAN INSTI-TUTE OF TECHNOLOGY GMBH	1115	1054	1111	1046	1050	1047	**1115**	1114	1112	1073

As shown in Table 4.11, the leading collaborators are predominantly higher education institutions, such as the Denmark University of Technology, Aarhus University, University of Gent, and University of Leuven; research centers, including Stichting Wageningen Research and Interuniversitair Micro-electronica Centrum; and innovative, technology-oriented enterprises, such as Fraunhofer-Gesellschaft zur Förderung der Angewandten Forschung e. V. (Fraunhofer) and the AIT Austrian Institute of

Technology GmbH. Across all methods, the number of links was generally underestimated, with the most accurate predictions provided by the RF approaches.

The highest observed DC value corresponded to Fraunhofer-Gesellschaft zur Förderung der Angewandten Forschung e. V. (Fraunhofer), a German organization with approximately 30,000 employees. Recognized as the largest provider of applied research and development services in Europe, Fraunhofer occupies a central position within the German research landscape, collaborating extensively with universities, the Max Planck Society, the Helmholtz Association of German Research Centers, the Gottfried Wilhelm Leibniz Scientific Association and the German Research Association. The headquarters is located in Munich, Germany. The remaining top collaborators are also closely linked to universities, with most non-university institutions ultimately owned or governed by academic entities.

4.2.6 Identification of Collaboration Communities

Collaboration communities are defined as sets of organizations where the observed collaborative ties exceed the predictions of the link estimation model. The optimal link prediction method (RF) underestimated collaborations by merely 1,618 instances across 1,198 organizations. The spatial distribution of these underestimated collaborations is visualized in Figure 4.10.

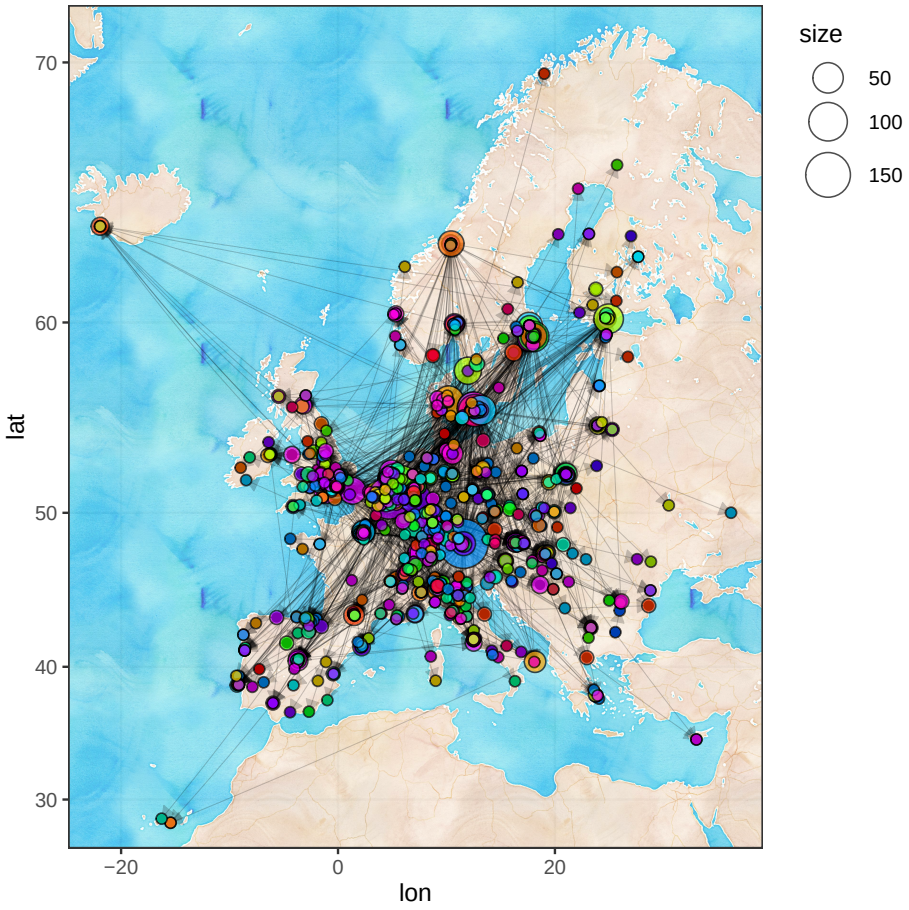


FIGURE 4.10: Economic collaboration communities. Nodes correspond to organizations, with edges representing underestimated collaborations between partners and coordinators. Node size scales with the logarithm of the logarithm of DC values. Modules are differentiated by color.

The identified communities comprised seven distinct components and 25 modules. As geographical distance was included as a predictor variable in the link estimation model, the module composition is geographically agnostic. However, Figure 4.10 reveals strong spatial clustering within the core EU nations, particularly Germany and the UK. The UK remains interconnected with continental Europe via hundreds of collaborative ties, a pattern likely to face disruption due to Brexit.

Module assignments (indicated by color) show no strict alignment with administrative boundaries. The largest modules (red) are predominantly concentrated in Germany and the United Kingdom. Communities in emerging EU member states exhibit greater heterogeneity and fewer underestimated collaborations.

The identified modules are referred in Table 4.12. The table shows the different modules identified by the number in column name Module, the number of entities relating to each module, and the affected countries from which the corresponding firms originate (please note that for UK the GB notation is used as official country code). There are big differences between the number of companies included in different modules. Although an investigation was conducted on the basis of the results shown in the table, a clear and comprehensive identification was not possible for the different modules. It was not successful in giving a proper explanation which would underline the clear and exact differentiation rule between the 25 modules

found. This indicates a new possible way for future research and investigation with probable participation of further information and data.

TABLE 4.12: Identified modules with the number of member firms and countries in them

Module	Member count	Countries
1	162	AT, BE, BG, CH, CZ, DE, DK, EE, ES, FI, FR, GB, HR, HU, IE, IS, IT, LT, LU, LV, MT, NL, NO, PL, PT, RO, SE, SI, SK
2	128	AT, BE, CH, DE, DK, ES, FI, FR, GB, HR, IE, IT, LT, LU, MT, NL, NO, PL, PT, RO, SE, SI, SK
3	108	AT, BE, BG, CH, CY, CZ, DE, DK, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MT, NL, NO, PL, PT, RO, SE, SI, SK
4	83	AT, BE, CH, CZ, DE, DK, ES, FI, FR, GB, HR, IE, IS, IT, LT, LU, MT, NL, NO, PL, PT, RO, SE, SI, SK
5	82	AT, BE, BG, CH, DE, DK, ES, FR, GB, HU, IS, IT, LT, LU, MT, NL, NO, PL, PT, RO, SE, SI, SK
6	69	AT, BE, CH, CZ, DE, DK, ES, FR, GB, GR, HU, IT, LT, LU, MT, NL, NO, PL, PT, RO, SE, SI, SK
7	69	AT, BE, CH, CZ, DE, DK, ES, FI, FR, GB, HU, IE, IS, IT, LT, LU, MT, NL, NO, PL, PT, RO, SE, SI, SK
8	67	AT, BE, BG, CZ, DE, DK, EE, FI, FR, GB, GR, HU, IE, IS, IT, LT, LU, LV, MT, NL, NO, PL, PT, RO, SE, SI, SK
9	58	AT, BE, CH, CZ, DE, EE, ES, FI, FR, GB, HU, IE, IS, IT, LT, LU, LV, MT, NL, NO, PL, PT, RO, SE, SI, SK
10	50	AT, BE, CH, DE, DK, ES, FI, FR, GB, HU, IE, IT, LT, LU, MT, NL, NO, PL, PT, RO, SE, SI, SK
11	50	AT, BG, CH, CY, CZ, DE, DK, ES, FI, FR, GB, HU, IE, IS, IT, LT, LU, LV, MT, NL, NO, PL, PT, RO, SE, SI, SK
12	48	AT, BE, CZ, DE, DK, ES, FR, GB, IE, IS, IT, NL, NO, PL, PT, RO, SE, SI, SK
13	44	BE, BG, CH, CZ, DE, ES, FR, GB, GR, IT, LU, NL, NO, PL, PT, RO, SE, SI, SK
14	39	AT, BE, BG, CH, DE, DK, ES, FI, FR, GB, HU, IE, IT, LT, LU, NL, NO, PL, PT, RO, SE, SI, SK
15	39	AT, BE, CH, DE, DK, FR, GB, HU, IE, IT, PL, SE
16	29	AT, BE, CH, DE, FR, GB, GR, IT, LT, NL, NO, RO
17	24	AT, BE, DE, ES, FI, FR, GB, IE, IT, NL, PT, SE
18	16	BE, BG, CZ, DE, FI, FR, GB, LT, NL, PT, SE
19	9	AT, ES, FR, GB, IT, NL, NO, PL
20	8	DK, ES, FI, HU, NL, PL
21	7	ES, FI, FR, HU, LT, UA
22	3	DE, ES, NO
23	2	BG
24	2	FR, PL
25	2	FR, NL

Chapter 5

Discussion

The establishment of subsidiaries can be interpreted as a form of investment in which a parent company initiates operations in a new geographic region or country. This process involves the transfer of technological assets to the new location and the generation of employment opportunities. Consequently, the analysis of such corporate networks is of significant importance. However, comprehensive coverage of economic and ownership relationships between firms remains underrepresented in existing databases.

The proposed annual gravity model has shown that subsidiary formations are driven by technological and economic disparities (Table 4.2), with capital flows observed from the more developed regions to less developed ones. Integration with network models has also revealed that such investments are predominantly domestic (Figs. 4.6-4.7). Despite efforts by the European Union to promote integration, administrative boundaries continue to exert substantial influence on ownership network structures (see the subfigures in Fig. 4.6), with minimal evolution observed during the study period (Fig. 4.7).

The proposed annual GEN models are identified as the most effective in explaining the formation dynamics of corporate ownership networks (Fig. 4.1; centrality estimates in Table 4.4). These models also accurately predict regions with the highest investment attractiveness (Table 4.5). Through the integration of GEN with module detection algorithms, economically coherent communities can be delineated and rationalized. This combined approach highlights the central role of core nations — including France, Germany, Great Britain, and the Benelux countries — in shaping ownership patterns (Figs. 4.3-4.6). In particular, the depth of integration of Great Britain within European economic networks (Fig. 4.5) underscores its pivotal position within these core communities (Fig. 4.6).

All analytical approaches confirm that the collaboration network under investigation is both extensive and highly structured. The H2020 collaboration network has been characterized as highly concentrated yet fragmented, with a limited number of intermediate connecting communities (see Table 4.6). This network structure has been shown to be conducive to the application of link prediction methodologies (Zhou et al., 2009). Almost half of the organizations that possess a Bvd Id number in Orbis database – and therefore access to corporate data – are classified as large or very large companies (see Table 4.7). These entities are considerably more capital intensive than small and medium-sized enterprises, whose support remains a core objective of H2020. Although Fig. 4.8 demonstrates the importance of all variables incorporated in the link prediction model, the findings indicate that organizations with substantial capital, as well as robust economic and technological backgrounds, display a higher propensity for collaboration (see Table 4.8; cf. Fig. 4.9). This tendency is even more pronounced among coordinators. Participation in multiple concurrent projects (as indicated by MULTI) or a record of previous successful projects

(as indicated by PROG) significantly enhances the likelihood of collaboration.

This outcome accounts for the high concentration, fragmentation, and emergence of small collaborative communities observed. Similar patterns were anticipated based on previous research Roediger-Schluga and Barber (2008) and Scherngell and Lata (2013); however, it has been established that the indicators MULTI and PROG, together with earned contribution (EC), exert a dominant influence on the evolution of collaborative relationships (see Fig. 4.9). These insights were made possible by utilizing a comprehensive dataset integrating multiple sources - Orbis, Cordis, Patstat, and Eurostat — thereby enabling the examination of novel variables, the identification of new relationships, and the achievement of high prediction accuracy.

Both generic approaches (such as LDA, QDA, and LogR) and non-generic black-box methods (such as SVM, RF, and XGBoost) have demonstrated the capacity to predict collaborative links. Although nongeneric methods achieved superior precision (see Table 4.9), interpretability was mainly supported by generic methods and the RF approach. The parameter tuning of non-generic methods only marginally improved the prediction performance (see Table 4.9; Table 4.10). The application of these methods resulted in notable improvements in predictive performance compared to previous studies on collaboration analysis (Chen et al., 2021; Wang et al., 2015a).

Analysis of current collaboration measures indicates that these partnerships exhibit a degree of predictability, suggesting a relatively fixed and stable network structure. Such stability can have both beneficial and adverse consequences. Predictable collaborations often reflect long-lasting and well-established partnerships, which can facilitate continuity in research and innovation activities. The ability to anticipate successful collaborations allows organizations to allocate resources efficiently and prioritize strengthening existing relationships. Predictable partnerships can also foster mutual trust, effective communication, shared goals, and positive outcomes. Conversely, entrenched partnerships may constrain the exploration of new ideas, technologies, or methodologies. Organizations that do not actively pursue diverse collaborations may miss opportunities for innovation and growth. Excess predictability can also discourage risk taking and exploration of new domains, thus impeding innovation and reducing organizational dynamism. Furthermore, a lack of diversity in perspectives, expertise, and resources may limit the ability to address complex challenges and stifle creativity.

The results further indicate that collaboration within the network is self-concentrating. The dominance of organizations with strong corporate, technological, or economic profiles is evident; however, in several instances, the actual number of connections is lower than predicted, resulting in a network that is more fragmented and concentrated than the predicted subgraph (see Table 4.10). This observation suggests the presence of additional undisclosed mechanisms that intensify concentration and fragmentation. In the case when all the investigated models underestimate the number of links in the networks, the most prominent and highly connected organizations still persist (see Table 4.11). These entities form collaboration communities that are geographically concentrated in Western Europe, specifically in Germany, Great Britain, Spain, and Italy (see Fig. 4.10). Such novel and significant findings are only discernible through the analysis of outliers, underscoring the importance of investigating weak signals in collaboration network analysis.

Chapter 6

Threats to Validity

The impact of threats to validity must be carefully considered throughout both the research process and the interpretation of the results. Although validity is an aspirational objective that cannot be guaranteed, the adoption of a structured approach from the literature, including conclusion, internal, construct and external validity Wohlin et al. (2012), enables the identification and mitigation of potential threats. In this dissertation, the analysis of validity will be addressed in detail in the following section.

Internal Validity refers to the extent to which a study can demonstrate a cause-and-effect relationship between the independent and dependent variables. Threats to internal validity often arise from factors such as selection bias, history, maturation, testing effects, instrumentation, and extraneous variables. Campbell's foundational work outlined specific threats to internal validity, which highlight the need for rigorous experimental controls to draw accurate inferences (Shadish, 2010).

In this work, during the construction of the data set for collaboration investigation, the members of the training and the test set were randomized, as mentioned in Section 3.3.2. The relationships between different variables were also examined and handled. For the ownership network, the complete data set was used, no subset was created, and only the missing values were excluded. With the number of executions of generic and nongeneric methods, the proposals from the relevant literature were used.

External Validity involves the generalization of research findings beyond specific study conditions to larger populations or different contexts. Factors that can threaten external validity include the selection of study participants, the specificity of the interventions used, and the environmental context in which the research is conducted.

During this work, only real-life databases were used, which contain data and information about real-world entities, and descriptive values were not changed, only used. Therefore, the information is not biased from any simulation result although due to the incompleteness of the databases some exclusion was made to include only relevant and consistent data into the examination. Therefore, the generalization of the usage of the proposed models and methods is possible as was also discussed earlier, and also the results can be considered as valid for real-world scenarios.

Construct Validity examines whether a study truly measures what it aims to measure. Threats include insufficient operational definitions and measurement tools that do not capture the intended constructs effectively. Research has revealed that a substantial number of studies do not discuss how their measures align with the underlying constructs of interest, raising concerns about their construct validity (Sjøberg and Bergersen, 2023). Poor operationalization can mislead research conclusions and skew findings, suggesting the need for a comprehensive assessment of construct validity throughout the research process (Lambert and Newman, 2022).

In this dissertation, the construction of the dataset, the investigation of the data, and also the methods used and applied were documented in Chapter 3. All methods and models used were inspected and the benefits as well as the weaknesses of them were considered and described. Several different methods and models were used and were transparently shown with the exact results and outcomes.
Conclusion Validity focuses on the proper application of statistical analyzes to derive conclusions from data. Issues such as low statistical power, incorrect use of statistical tests, and misinterpretation of results pose significant threats to conclusion validity. It has been established that many studies do not adequately address the validity of statistical conclusions, relying on flawed analytical approaches that yield incorrect conclusions.

In this work, all statistical significance was checked and properly documented and, where the desired significance level was not achieved, the possible reasons were identified and discussed. Conclusions were peer-reviewed with experts in the areas and also compared with the results of the already available literature. The tests and methods used were also verified and, as was referred to earlier, are widely used in the area of research.

Chapter 7

Summary and Conclusion

This dissertation, on the one hand, is focused on the integration of network analysis with descriptive and economic factors within explanatory frameworks, leveraging the strengths of diverse methodological approaches. A generalized yearly gravity-based economic null GEN model is proposed to predict the spatial structure of corporate ownership networks. Compared to the models introduced by Newman and Girvan (2004) and Expert et al. (2011), gravity-based approaches are shown to produce better estimates of global network properties, including centrality metrics. Furthermore, a gravity-driven modularity measure is introduced to delineate economically coherent communities, while the annualized formulation of the model enables the examination of temporal evolution in these communities.

The GEN model is positioned as a robust tool to advance the understanding of the mechanisms of network formation.

On the other hand, in this dissertation, an accurate link prediction model has been proposed for the analysis and prediction of H2020 collaborations, achieved by integrating multiple databases and expanding the set of variables analyzed with additional factors. Several contributions and implications have been established in this work, as outlined below.

In this work a highly accurate link prediction model has been developed. Two generic and three nongeneric methods were examined, with the nongeneric approaches found to yield superior accuracy, thereby indicating the significance of nonlinear relationships within the models. For decision makers, the achievable model accuracy has been demonstrated, which is advantageous to forecast and facilitate the development of future consortia in Framework Programmes FPs. The findings of this dissertation have supported effective model selection during the analytical process.

In addition, the principal factors influencing organizational collaboration have been identified and ranked. Embeddedness in prior projects, reflected in multi-project and program characteristics, as well as earned contributions, has been shown to be a primary determinant of the likelihood of collaboration. These insights can assist decision-makers in effectively supporting the formation of future FP collaborations.

Finally, in this study, outlier collaboration communities have been specified, predominantly comprising organizations from EU core countries. The identification of these weak, yet significant, observations highlights the need for a more in-depth analysis of collaboration patterns among core countries as a prospective research direction. This finding offers stakeholders the opportunity to examine collaboration concentrations within FPs and to either reinforce existing collaborations or promote greater balance by engaging organizations outside the core countries.

7.1 Research Theses

According to the research questions asked in Section 1.5 three Research theses were formulated that considered the results of Section 4, 5 and 6.

RT1: In this dissertation, it has been demonstrated that the gravity-based economic null model reduces link prediction error in comparison to the Newman-Girvan and Expert models, while also enabling the identification of economic-investment communities and providing superior estimation of derived network parameters relative to the aforementioned models. The model has indicated that company establishments within the European Union are influenced by technological and economic disparities, with investment flows directed from more developed regions toward less developed ones.

RT2: The gravity-based economic null model proposed in this work has revealed that primary investment flows are strongly shaped by administrative boundaries (country borders), as investments are predominantly established within national borders, resulting in capital largely remaining domestic. Although the European Union seeks to enhance integration and promote economic equality across all member states, administrative boundaries continue to exert a significant effect on the formation of the European ownership network.

RT2.1: The proposed annual model introduced in this study, has the capability to identify temporal changes in economic-investment communities.

RT3: The proposed model for the collaboration network analysis together with the integrated machine learning techniques have been shown in this dissertation to outperform benchmark models and further improve predictive accuracy. Relationships among 23 validated corporate and economic predictors were captured, and all defined variables were identified as important, underscoring the advantages of employing a comprehensive dataset. The proposed model has also revealed that the most influential variables for collaboration formation pertain to previous successful collaborations within the same framework programme. Outlier collaboration communities, restricted to the European Unions core countries, were identified.

RT3.1: In regards of link prediction in European Union's Framework Programmes collaboration networks, the non-generic methods perform better than the generic ones, indicating nonlinear relationships in the model. The best performing methods among the applied non-generic ones are the Random Forest based ones.

Table 7.1 summarizes the research with Research question, assumptions and theses together.

7.2 Implications

The findings of the present study hold significant relevance for decision-makers and policy formulators within the investigated domains. The following implications are derived from the results and are intended to inform strategic planning and policy development.

The analysis of the ownership network revealed that the predominant direction of investment flows is from economically advanced regions toward less developed areas. This insight provides a basis for governmental authorities to prioritize national infrastructure development initiatives in underdeveloped regions, thereby enhancing their attractiveness for future subsidiary establishment. Such targeted interventions may contribute to reducing regional disparities and fostering balanced economic growth.

Furthermore, these findings are of particular importance to policymakers within the European Union. When designing supportive programs aimed at mitigating territorial inequalities, the results suggest that directives could be refined to focus on specific geographic units, including NUTS 3 regions, thereby enabling more precise allocation of resources and tailored regional development strategies.

The temporal stability observed in the ownership network between 2010 and 2018 indicates that the principal economic-investment communities have remained largely unchanged over this period. This persistence implies that investment flows predominantly circulate within the core EU countries in the long term. However, this equilibrium may be altered through deliberate policy interventions, such as the implementation of tax incentives or regulatory adjustments, which could redirect investment towards targeted regions.

The investigation of the H2020 collaboration network also yields several policy-relevant implications. The observed scale-free nature of link formation suggests that entities with established collaborative histories possess a disproportionately higher likelihood of securing additional partnerships compared to newcomers. In response, FPs policymakers might consider instituting upper limits on the number of projects awarded to individual entities, potentially differentiated by organizational size, to promote equitable participation.

Additionally, specific provisions could be introduced to facilitate the engagement of entities with no prior participation in subsidized FPs projects. By easing eligibility criteria or providing targeted support, these measures could incentivize broader involvement and diversification within funding programs.

Finally, the feature ranking derived from this study offers a valuable tool for balancing future FPs consortia formations. This information, combined with the outcomes of outlier analyses, can assist policymakers in modulating the concentration of collaborative connections, thereby optimizing the distribution of resources and enhancing the inclusivity and effectiveness of funded projects.

Collectively, these implications underscore the potential for evidence-based policy design informed by network analysis to foster more balanced and dynamic economic and innovation systems.

7.3 Contribution to the literature

The proposed GEN model (published in Kosztyán et al. (2022b)) provided a better link prediction and revealed information about the subsidiary structure within the European Union, which showed some differences from the goal of economic equality which is an important goal of the EU. The EICs also identified a novelty in the literature that shed light on the areas to further develop.

As a result of the collaboration investigation (published in Kosztyán et al. (2024)), an accurate model was developed for link prediction, including corporate, economic, patent and collaboration databases which are connected to each other to enable

complex research. The evaluation provided of the models supports researchers to achieve efficient model selection in the collaboration analysis field.

Key economic, technological and collaboration factors were also identified and showed that they influence the collaboration between entities.

The proposed model is capable of specifying the communities which are considered as outliers from the collaboration communities point of view in regards to the collaboration connections being denser than the predicted values.

TABLE 7.1: Summary table for Research Questions, Assumptions and Theses

Item	Statement
RQ1:	Can the proposed gravity-based economic null model improve link prediction and network coefficient estimation, identify stable Emergent Innovation Communities, and provide insights into their spatial and temporal dynamics?
RA1:	Gravity-driven economic principles dominate ownership network formation, with gravity-based economic null model predictions reflecting real-world investment flows more accurately than topology-only models.
RT1:	In this dissertation, it has been demonstrated that the gravity-based economic null model reduces link prediction error in comparison to the Newman-Girvan and Expert models, while also enabling the identification of economic-investment communities and providing superior estimation of derived network parameters relative to the aforementioned models. The model has indicated that company establishments within the European Union are influenced by technological and economic disparities, with investment flows directed from more developed regions toward less developed ones.
RQ2:	To what extent do administrative borders influence investment flows, and how do these effects change when controlling for geographical distance?
RA2:	Administrative borders create structural breaks in ownership networks independent of geographic proximity, persisting across temporal layers.
RT2:	The gravity-based economic null model proposed in this work has revealed that primary investment flows are strongly shaped by administrative boundaries (country borders), as investments are predominantly established within national borders, resulting in capital largely remaining domestic. Although the European Union seeks to enhance integration and promote economic equality across all member states, administrative boundaries continue to exert a significant effect on the formation of the European ownership network.
RT2.1:	The proposed annual model introduced in this study, has the capability to identify temporal changes in economic-investment communities.
RQ3:	Can the proposed model, applied to a comprehensive dataset, enhance our understanding and predictive accuracy of organizational collaboration and community structures in Framework Programmes beyond current benchmarks?
RA3:	Machine learning techniques including generic and non-generic approaches can be beneficial for improving the prediction of the connections in the collaboration network of the Horizon 2020 Programme and with the proper model, the influential factors can also be identified.

Item	Statement
RT3:	The proposed model for the collaboration network analysis together with the integrated machine learning techniques have been shown in this dissertation to outperform benchmark models and further improve predictive accuracy. Relationships among 23 validated corporate and economic predictors were captured, and all defined variables were identified as important, underscoring the advantages of employing a comprehensive dataset. The proposed model has also revealed that the most influential variables for collaboration formation pertain to previous successful collaborations within the same framework programme. Outlier collaboration communities, restricted to the European Unions core countries, were identified.
RT3.1:	In regards of link prediction in European Union's Framework Programmes collaboration networks, the non-generic methods perform better than the generic ones, indicating nonlinear relationships in the model. The best performing methods among the applied non-generic ones are the Random Forest based ones.

Chapter 8

Limitations and Future Research

In this dissertation, the analysis focused on European organizations, although corporate, patent, and GDP data are globally accessible. However, the interpretation of NUTS 3 regions remains confined to the European context. This work has demonstrated that Great Britain, particularly England, is interconnected with the European Union through extensive collaborative ties. The examination of post-Brexit network dynamics is identified as a compelling avenue for future research.

Within the proposed GEN model, a yearly estimation framework is being developed for multilayer ownership networks. Furthermore, the gravity model framework can be extended to accommodate multiplex networks, where multiple interrelated networks are analyzed concurrently. Finally, the inclusion of industry-level analysis in ownership structures is believed to yield deeper insights into the formation of parent-subsidiary relationships.

Although the created research database is unique from the perspective of how many different and heterogeneous sources were used and combined, there are still possibilities to further extend it with additional information such as tax-environment-relevant information related to areas, tariff information or cultural information. The database can also be extended to further territories not to focus only on European entities.

The collaboration was investigated based on the H2020 Framework Programme, although other programs can also be considered, and a comparative study would also be beneficial. As Horizon Europe [1] already started, the same research can be repeated after the first projects are closed within this funding time frame and the differences (if any) can be identified.

The large and very large companies are overrepresented in the data sources used, especially in the Amadeus and Orbis data sets. This biased database could be extended with more accurate information for smaller entities in the future although until today there was not identified any reliable data source which contains the needed company-level information for these kind of entities in a structured way.

From a database perspective, there are also possibilities to further extend it with information about the classification of economic activities which can be derived from 4 digit NACE codes, where NACE refers to the Nomenclature générale des Activités économiques dans les Communautés Européennes, which translates to Statistical classification of economic activities in the European Community. With this kind of extension, it would be possible to differentiate the entities in both investigation fields and identify possible differences in the network based on the fields in which the different companies are active. This NACE information is nowadays available in different data sources so it's integration into the already created research database is not too hard.

[1] https://research-and-innovation.ec.europa.eu/funding/funding-opportunities/funding-programmes-and-open-calls/horizon-europe_en/

Another improvement can be to integrate tax related information into the database, which most probably can be defined on national level. It would be interesting to see the results from the same analysis whether any differences occur in the formation of the networks in case we also consider different amounts of taxes in different areas.

With the above mentioned extensions, the outlier analysis in the collaboration network becomes also possible to be redone and having some more detailed information, the explanation of the different modules could be feasible.

Appendix A

Appendix A

The indicators described in Table 3.1 - Applied indicators for ownership network - are defined as follows:

Indicator	Name	Unit	Description / Formula
m_1	Total assets	Thousand €	Total assets = fixed assets + current assets
m_2	Solvency ratio (asset based)	%	(Shareholder funds / Total assets) × 100
m_3	Shareholders' funds	Thousand €	Total equity (capital + other shareholders' funds)
m_4	ROE using P/L before tax	%	(Profit before tax / Shareholder funds) × 100
m_5	ROCE using P/L before tax	%	(Profit before tax + Interest paid) / (Shareholder funds + Noncurrent liabilities) × 100
m_6	Profit margin	%	(Profit before tax / Operating revenue) × 100
m_7	P/L for period	Thousand €	Net income
m_8	P/L before tax	Thousand €	Operating profit + Financial profit
m_9	Operating revenue (turnover)	Thousand €	Net sales + Other operating revenues + Stock variations (excluding VAT)
m_{10}	Fixed assets	Thousand €	Intangible + Tangible + Other fixed assets (after depreciation)
m_{11}	Number of employees	Count	Total number of employees
m_{12}	Current ratio	-	Current assets / Current liabilities
m_{13}	Cash flow	Thousand €	Profit for period + Depreciation
m_{14}	Number of companies	Count	Number of companies in the NUTS3 region
m_{15}	GDP per capita in PPP	Thousand €	Compares economic productivity and standard of living using PPP methodology
m_{16}	Patents	Count	Number of patent applications and granted patents

TABLE A.1: Financial and economic indicators

Appendix B

Appendix B

TABLE B.1: Summary table of complete gravity models with regression coefficients and absolute errors of the estimated centralities

Coefficients	2010 β		2011 β		2012 β		2013 β		2014 β		2015 β		2016 β		2017 β		2018 β	
(Intercept)	1.5886	***	1.5838	***	1.5350	***	1.5797	***	1.5889	***	1.5754	***	1.2931	***	2.0835	***	2.1312	***
$D_{i,j}$	-0.4730	***	-0.4731	***	-0.4734	***	-0.4741	***	-0.4759	***	-0.4739	***	-0.4740	***	-0.4729	***	-0.4737	***
TA_i	-0.0510	***	-0.0515	***	-0.0900	***	-0.1057	***	-0.0856	***	-0.1212	***	-0.1363	***	-0.1592	***	-0.1773	***
SR_i	-0.1107	***	-0.1053	***	-0.0792	***	-0.1055	***	-0.1252	***	-0.1277	***	-0.1572	***	-0.1548	***	-0.1538	***
SH_i	-0.0227	*	-0.0228		-0.0265	*	0.0079		-0.0207		0.0073		0.0301	*	0.0265	*	0.0678	***
RB_i	-0.1127	***	-0.0825	***	-0.1015	***	-0.0978	***	-0.1223	***	-0.1166	***	-0.1453	***	-0.1253	***	-0.1268	***
RCB_i	-0.0114		-0.0066		-0.0237	***	-0.0244	***	-0.0262	***	-0.0176	***	-0.0160	**	-0.0174	**	-0.0151	**
PM_i	0.0937	***	0.0514	***	0.0840	***	0.0980	***	0.0932	***	0.1134	***	0.1312	***	0.1120	***	0.1141	***
PLF_i	-0.0331	***	-0.0412	***	-0.0418	***	-0.0925	***	-0.0591	***	-0.0965	***	-0.0472	***	-0.0460	***	-0.0355	**
PLB_i	0.0115		0.0209		0.0081		0.0451	***	0.0305	*	0.0448	**	-0.0184		0.0426	**	0.0090	
OR_i	0.0269	***	0.0223	**	0.0381	***	0.0283	***	0.0190	*	0.0374	***	0.0794	***	0.0476	***	0.0256	**
FA_i	0.0838	***	0.0879	***	0.1118	***	0.1173	***	0.1129	***	0.1173	***	0.1068	***	0.1019	***	0.1046	***
EN_i	0.0005		0.0070		0.0137	***	0.0240	***	0.0449	***	0.0420	***	0.0117	*	0.0020		0.0075	
CR_i	0.1097	***	0.1041	***	0.0853	***	0.1207	***	0.1483	***	0.1034	***	0.1724	***	0.0986	***	0.0644	***
CF_i	0.0064		-0.0014		0.0117	*	0.0042		-0.0141	**	0.0039		0.0054		0.0087		0.0050	
CO_i	0.2042	***	0.2058	***	0.2085	***	0.2082	***	0.2133	***	0.2184	***	0.2159	***	0.2145	***	0.2074	***
GDP_i	-0.0014		-0.0011		-0.0001		-0.0004		-0.0004		-0.0222	***	-0.0131	**	-0.0157	**	0.0214	*
PI_i	0.0028		0.0031	*	0.0019		0.0007		0.0020		0.0012		0.0026		0.0041	**	0.0127	***
TA_j	-0.0264	*	-0.0400	**	-0.0267	*	0.0218		0.0543	***	0.0577	***	0.0302	*	0.0242		-0.0058	
SR_j	-0.0429	***	-0.0754	***	-0.0523	***	-0.0604	***	-0.0200		0.0386	*	0.0139		-0.0296		-0.0567	***
SH_j	0.0777	***	0.0961	***	0.0744	***	0.0740	***	0.0673	***	0.0372	**	0.0629	***	0.0579	***	0.0693	***
RB_j	-0.0252	**	-0.0240	**	-0.0408	***	-0.0468	***	-0.0475	***	-0.0550	***	-0.0694	***	-0.0826	***	-0.0777	***
RCB_j	-0.0152	**	-0.0130	*	-0.0222	***	-0.0172	***	-0.0138	**	0.0044		0.0032		-0.0069		-0.0028	
PM_j	0.0566	***	0.0638	***	0.0578	***	0.0517	***	0.0237	*	0.0249	*	0.0482	***	0.0608	***	0.0529	***
PLF_j	-0.0192	*	-0.0374	***	-0.0497	***	-0.0620	***	0.0321	**	0.0034		0.0287	*	0.0208		0.0584	***
PLB_j	0.0026		-0.0072		0.0055		0.0147		-0.0954	***	-0.0564	***	-0.1063	***	-0.0778	***	-0.1027	***
OR_j	0.0240	***	0.0407	***	0.0403	***	0.0098		0.0206	*	0.0314	***	0.0729	***	0.0500	***	0.0374	***
FA_j	-0.0369	***	-0.0338	***	-0.0284	***	-0.0414	***	-0.0535	***	-0.0385	***	-0.0376	***	-0.0348	***	-0.0286	***
EN_j	-0.0359	***	-0.0246	***	-0.0262	***	-0.0135	**	-0.0125	**	-0.0194	***	-0.0466	***	-0.0536	***	-0.0483	***
CR_j	0.0542	***	0.0805	***	0.0655	***	0.0807	***	0.0946	***	0.0231		0.0333	*	-0.0411	**	-0.0239	
CF_j	-0.0067		-0.0104	*	-0.0064		-0.0199	***	-0.0308	***	-0.0342	***	-0.0296	***	-0.0193	***	-0.0144	***
CO_j	0.2152	***	0.2139	***	0.2168	***	0.2182	***	0.2218	***	0.2280	***	0.2222	***	0.2222	***	0.2202	***
GDP_j	-0.0085	***	-0.0091	***	-0.0089	***	-0.0099	***	-0.0093	***	-0.0163	**	-0.0059		-0.0162	**	-0.0157	*
PI_j	-0.0067	***	-0.0054	***	-0.0051	***	-0.0073	***	-0.0072	***	-0.0106	***	-0.0083	***	-0.0050	**	-0.0007	
Adj. R^2	0.4061	***	0.4057	***	0.4062	***	0.4073	***	0.4087	***	0.4072	***	0.4082	***	0.4058	***	0.4067	***
ϵ^{grav}	0.0078		0.0073		0.0078		0.0077		0.0076		0.0080		0.0081		0.0080		0.0082	
$\epsilon_{C_D^+}$	3.0414		5.7021		3.3355		3.5763		4.2312		3.8885		4.6273		6.5212		4.7929	
$\epsilon_{C_D^-}$	4.5173		3.6933		5.1694		4.9229		4.9568		5.5349		4.1927		6.3334		5.5961	
ϵ_{C_B}	142.7669		183.9388		139.3517		160.2476		161.9900		147.5215		167.2923		185.0560		182.4788	
$\epsilon_{C_C^+}$	2.72E-06		3.25E-06		2.46E-06		2.22E-06		1.75E-06		2.77E-06		2.99E-06		1.91E-06		2.43E-06	
$\epsilon_{C_C^-}$	4.62E-06		1.97E-06		4.71E-06		4.06E-06		3.60E-06		4.59E-06		2.77E-06		2.26E-06		3.77E-06	
ϵ_{C_H}	1.98E-05		1.67E-05		2.06E-05		1.94E-05		1.85E-05		2.13E-05		1.81E-05		1.64E-05		2.06E-05	
ϵ_{C_A}	1.42E-05		1.93E-05		1.28E-05		1.38E-05		1.39E-05		1.47E-05		1.45E-05		1.22E-05		1.53E-05	
ϵ_{C_P}	2.83E-05		3.45E-05		2.62E-05		3.15E-05		3.18E-05		3.25E-05		2.23E-05		2.05E-05		3.00E-05	

Values are significant at: * p=0.05, ** p=0.01, *** p=0.001 levels.

Bibliography

Abonyi, János, Czvetkó, Tímea, and Honti, Gergely Marcell (2020). *Are Regions Prepared for Industry 4.0?: The Industry 4.0+ Indicator System for Assessment*. Springer Nature.

Abrham, Josef and Vosta, Milan (Sept. 2011). *Regional differentiation, agglomeration and clusters within the EU*. ERSA conference papers ersa10p1660. European Regional Science Association. URL: https://ideas.repec.org/p/wiw/wiwrsa/ersa10p1660.html.

Adomako, Samuel, Amankwah-Amoah, Joseph, Debrah, Yaw A., Khan, Zaheer, Chu, Irene, and Robinson, Catherine (2020). "Institutional Voids, Economic Adversity and Inter-firm Cooperation in an Emerging Market: The Mediating Role of Government R&D Support". In: *British Journal of Management*. DOI: 10.1111/1467-8551.12443.

Ahmad, Iftikhar, Akhtar, Muhammad Usman, Noor, Salma, and Shahnaz, Ambreen (2020). "Missing Link Prediction Using Common Neighbor and Centrality Based Parameterized Algorithm". In: *Scientific Reports*. DOI: 10.1038/s41598-019-57304-y.

Ahmad, Iftikhar, Basheri, Mohammad, Iqbal, Muhammad Javed, and Rahim, Aneel (2018). "Performance comparison of support vector machine, random forest, and extreme learning machine for intrusion detection". In: *IEEE access* 6, pp. 33789–33795. DOI: https://doi.org/10.1109/ACCESS.2018.2841987.

Albert, Réka and Barabási, Albert-László (2002). "Statistical mechanics of complex networks". In: *Reviews of Modern Physics* 74.1, p. 47.

Albert, Réka and Barabási, Albert-László (2002). "Statistical Mechanics of Complex Networks". In: *Reviews of Modern Physics*. DOI: 10.1103/revmodphys.74.47.

Ali, Waris, Frynas, Jêdrzej George, and Wilson, Jeffrey (2024). "Corporate–NGO Collaboration and CSR Disclosure – The Moderating Role of Corporate Profitability". In: *Journal of Applied Accounting Research*. DOI: 10.1108/jaar-08-2023-0238.

Alibrahim, Hussain and Ludwig, Simone A (2021). "Hyperparameter optimization: Comparing genetic algorithm against grid search and bayesian optimization". In: *2021 IEEE Congress on Evolutionary Computation (CEC)*. IEEE, pp. 1551–1559. DOI: https://doi.org/10.1109/CEC45853.2021.9504761.

Amini, Hamed, Cont, Rama, and Minca, Andreea (2013). "Systemic risk and central counterparty clearing". In: *Journal of Financial Stability* 9.4, pp. 585–603.

Anderson, James E (2011). "The gravity model". In: *Annual Review of Economics* 3.1, pp. 133–160.

Andrade Rojas, Mariana Giovanna, Ramírez-Solís, Edgar Rogelio, and Zhu, Jianjun (2018). "Innovation and Network Multiplexity: R&D and the Concurrent Effects of Two Collaboration Networks in an Emerging Economy". In: *Research Policy*. DOI: 10.1016/j.respol.2018.03.018.

Annette, Lucy (2021). "The Development of Horizon Europe". In: *Impact*. DOI: 10.21820/23987073.2021.4.30.

Annette, Lucy (2025). "Taking Stock: Reflections on Horizon 2020". In: *Impact*. DOI: 10.21820/23987073.2025.2.63.

Arenas, Alex, Fernandez, Alberto, and Gomez, Sergio (2008). "Analysis of the structure of complex networks at different resolution levels". In: *New Journal of Physics* 10.5, p. 053039.

Artzy-Randrup, Yael, Fleishman, Sarel J, Ben-Tal, Nir, and Stone, Lewi (2005). "Comment on "Network motifs: simple building blocks of complex networks" and "Superfamilies of evolved and designed networks"". In: *Science* 305.5687, pp. 1107–1107.

Asero, Vincenzo, Gozzo, Simona, and Tomaselli, Venera (2016). "Building tourism networks through tourist mobility". In: *Journal of Travel Research* 55.6, pp. 751–763.

Babić, Milan, Garcia-Bernardo, Javier, and Heemskerk, Eelke M. (2019). "The Rise of Transnational State Capital: State-Led Foreign Investment in the 21st Century". In: *Review of International Political Economy*. DOI: 10.1080/09692290.2019.1665084.

Baier, Scott L and Bergstrand, Jeffrey H (2007). "Do free trade agreements actually increase members' international trade?" In: *Journal of International Economics* 71.1, pp. 72–95.

Barabási, Albert-László and Albert, Réka (1999a). "Emergence of scaling in random networks". In: *Science* 286.5439, pp. 509–512. DOI: 10.1126/science.286.5439.509.

Barabási, Albert-László and Albert, Réka (1999b). "Emergence of scaling in random networks". In: *Science* 286.5439, pp. 509–512.

Barabási, Albert-László (2016). *A hálózatok tudománya*. Libri, pp. 135–138.

Barbu, Giorgiana-Raluca and Niţă, Mihai Răzvan (2025). "Nature-Based Solutions for Climate Resilience in EU R&I Framework Programmes Horizon 2020 and Horizon Europe". In: DOI: 10.5194/egusphere-egu24-19213.

Barthélemy, Marc (2011). "Spatial networks". In: *Physics Reports* 499.1–3, 1—101. ISSN: 03701573. DOI: 10.1016/j.physrep.2010.11.002.

Bastami, Esmaeil, Mahabadi, Aminollah, and Taghizadeh, Elias (2019). "A gravitation-based link prediction approach in social networks". In: *Swarm and evolutionary computation* 44, pp. 176–186. DOI: https://doi.org/10.1016/j.swevo.2018.03.001.

Batty, Michael (2008). "The Size, Scale, and Shape of Cities". In: *Science*. DOI: 10.1126/science.1151419.

Bavelas, Alex (1950). "Communication patterns in task-oriented groups". In: *The journal of the acoustical society of America* 22.6, pp. 725–730.

Bayrak, Ahmet Engin and Polat, Faruk (2018). "Effective Feature Reduction for Link Prediction in Location-Based Social Networks". In: *Journal of Information Science*. DOI: 10.1177/0165551518808200.

Bergstra, James and Bengio, Yoshua (2012). "Random search for hyper-parameter optimization." In: *Journal of machine learning research* 13.2. URL: https://www.jmlr.org/papers/volume13/bergstra12a/bergstra12a.pdf?source=post_page.

Berry, Randall A. and Johari, Ramesh (2011). "Economic Modeling in Networking: A Primer". In: *Foundations and Trends® in Networking*. DOI: 10.1561/1300000011.

Bhattacharya, Kunal, Mukherjee, Gautam, Saramäki, Jari, Kaski, Kimmo, and Manna, Subhrangshu S (2008). "The international trade network: weighted network analysis and modelling". In: *Journal of Statistical Mechanics: Theory and Experiment* 2008.2, P02002.

Bi, Yilin, Jiao, Xinshan, Lee, Yan-Li, and Zhou, Tao (2024). "Inconsistency Among Evaluation Metrics in Link Prediction". In: *Pnas Nexus*. DOI: 10.1093/pnasnexus/pgae498.

Bickley, Steve J., Chan, Ho Fai, and Torgler, Benno (2022). "Artificial Intelligence in the Field of Economics". In: *Scientometrics*. DOI: 10.1007/s11192-022-04294-w.

Binte Azhar, Nurul Asyikeen, Pan, Gary, Sun, Seow Poh, Koh, Andrew, and Tay, Wan Ying (2019). "Text Analytics Approach to Examining Corporate Social Responsibility". In: *Asian Journal of Accounting and Governance*. DOI: 10.17576/ajag-2019-11-08.

Blažek, Roman, Durana, Pavol, Michulek, Jakub, and Blazekova, Kristina (Mar. 2023). "Does the Size of the Business Still Matter, or Is Profitability under New Management, by Order of the COVID-19?" In: *Journal of Risk and Financial Management* 16, p. 219. DOI: 10.3390/jrfm16040219.

Blondel, Vincent D, Guillaume, Jean-Loup, Lambiotte, Renaud, and Lefebvre, Etienne (2008). "Fast unfolding of communities in large networks". In: *Journal of Statistical Mechanics: Theory and Experiment* 2008.10, P10008. DOI: 10.1088/1742-5468/2008/10/p10008. URL: https://doi.org/10.1088/1742-5468/2008/10/p10008.

Bollobás, Béla, Riordan, Oliver, Spencer, Joel, and Tusnády, Gábor (2011). "The Degree Sequence of a Scale-Free Random Graph Process". In: *Complex Networks*. Princeton University Press, pp. 384–399. DOI: 10.1515/9781400841356.384.

Boss, Michael, Elsinger, Helmut, Summer, Martin, and Thurner, Stefan (2004). "Network topology of the interbank market". In: *Quantitative Finance* 4.6, pp. 677–684.

Burger, Martijn, Oort, Frank van, and Linders, Gert-Jan (2009). "On the Specification of the Gravity Model of Trade: Zeros, Excess Zeros and Zero-inflated Estimation". In: *Spatial Economic Analysis* 4.2, pp. 167–190. DOI: 10.1080/17421770902834327. eprint: https://doi.org/10.1080/17421770902834327. URL: https://doi.org/10.1080/17421770902834327.

Cai, Yuzhuo (2023). "Towards a new model of EU-China innovation cooperation: Bridging missing links between international university collaboration and international industry collaboration". In: *Technovation* 119, p. 102553. DOI: https://doi.org/10.1016/j.technovation.2022.102553.

Chang, Han-wen and Huang, Mu-Hsuan (2013). "Prominent institutions in international collaboration network in astronomy and astrophysics". In: *Scientometrics* 97, pp. 443–460. DOI: https://doi.org/10.1007/s11192-013-0976-x.

Chen, Chen, He, Jingrui, Bliss, Nadya, and Tong, Hanghang (2017a). "Towards Optimal Connectivity on Multi-Layered Networks". In: *Ieee Transactions on Knowledge and Data Engineering*. DOI: 10.1109/tkde.2017.2719026.

Chen, Chen, Tong, Hanghang, Xie, Lei, Ying, Lei, and He, Qing (2017b). "Cross-Dependency Inference in Multi-Layered Networks". In: *Acm Transactions on Knowledge Discovery From Data*. DOI: 10.1145/3056562.

Chen, Juntao and Zhu, Quanyan (2016). "Resilient and decentralized control of multi-level cooperative mobile networks to maintain connectivity under adversarial environment". In: *2016 IEEE 55th Conference on Decision and Control (CDC)*, pp. 5183–5188. DOI: 10.1109/CDC.2016.7799062.

Chen, Wei, Qu, Hui, and Chi, Kuo (2021). "Partner selection in China interorganizational patent cooperation network based on link prediction approaches". In: *Sustainability* 13.2, p. 1003. DOI: https://doi.org/10.3390/su13021003.

Chesbrough, Henry William (2003). *Open innovation: The new imperative for creating and profiting from technology*. Harvard Business Press. DOI: https://doi.org/10.1016/j.jengtecman.2004.05.003.

Chi, Kuo, Yin, Guisheng, Dong, Yuxin, and Dong, Hongbin (2019). "Link prediction in dynamic networks based on the attraction force between nodes". In: *Knowledge-Based Systems* 181, p. 104792. DOI: https://doi.org/10.1016/j.knosys.2019.05.035.

Chuanming, Yu, Yutian, Gong, Xiaoli, Zhao, and Lu, An (2017). "Collaboration recommendation of finance research based on multi-feature fusion". In: *Data Analysis and Knowledge Discovery* 1.8, pp. 39–47. DOI: https://doi.org/10.11925/infotech.2096-3467.2017.08.05.

Chávez-Bustamante, Felipe, Mardones-Arias, Elliott, Rojas-Mora, Julio, and Tijmes, Jaime (2023). "A Forgotten Effects Approach to the Analysis of Complex Economic Systems: Identifying Indirect Effects on Trade Networks". In: *Mathematics*. DOI: 10.3390/math11030531.

Cséfalvay, Zoltán and Gkotsis, Petros (2022). "Robotisation race in Europe: the robotisation chain approach". In: *Economics of Innovation and New Technology* 31.8, pp. 693–710. DOI: https://doi.org/10.1080/10438599.2020.1849968.

Czvetkó, Tímea, Honti, Gergely, and Abonyi, János (2021). "Regional development potentials of Industry 4.0: Open data indicators of the Industry 4.0+ model". In: *Plos one* 16.4, e0250247.

Dahesh, Mehran Badin, Tabarsa, Gholamali, Zandieh, Mostafa, and Hamidizadeh, Mohammadreza (2020). "Reviewing the intellectual structure and evolution of the innovation systems approach: A social network analysis". In: *Technology in Society* 63, p. 101399.

Darko, Josephine, Aribi, Zakaria Ali, and Uzonwanne, Godfrey (2016). "Corporate Governance: The Impact of Director and Board Structure, Ownership Structure and Corporate Control on the Performance of Listed Companies on the Ghana Stock Exchange". In: *Corporate Governance*. DOI: 10.1108/cg-11-2014-0133.

Daudin, Jean-Jacques, Picard, Franck, and Robin, Stéphane (2007). "A Mixture Model for Random Graphs". In: *Statistics and Computing*. DOI: 10.1007/s11222-007-9046-7.

De Noni, Ivan, Orsi, Luigi, and Belussi, Fiorenza (2018). "The role of collaborative networks in supporting the innovation performances of lagging-behind European regions". In: *Research Policy* 47.1, pp. 1–13. DOI: https://doi.org/10.1016/j.respol.2017.09.006.

De Prato, Giuditta and Nepelski, Daniel (2014). "Global technological collaboration network: Network analysis of international co-inventions". In: *The Journal of Technology Transfer* 39.3, pp. 358–375. DOI: https://doi.org/10.1007/s10961-012-9285-4.

de Sola Pool, Ithiel and Kochen, Manfred (1978). "Contacts and influence". In: *Social Networks* 1.1, pp. 5–51. ISSN: 0378-8733. DOI: https://doi.org/10.1016/0378-8733(78)90011-4. URL: https://www.sciencedirect.com/science/article/pii/0378873378900114.

Demir, Ferhat and Lukeš, Martin (2024). "Collaboration of Corporates With Coworking Spaces: Different Pathways to Develop Innovation Capabilities". In: *R and D Management*. DOI: 10.1111/radm.12697.

Demir, Selçuk and Şahin, Emrehan Kutluğ (2022). "Liquefaction prediction with robust machine learning algorithms (SVM, RF, and XGBoost) supported by genetic algorithm-based feature selection and parameter optimization from the perspective of data processing". In: *Environmental Earth Sciences* 81.18, p. 459. DOI: https://doi.org/10.1007/s12665-022-10578-4.

Deng, Xiaolong, Sun, Jufeng, and Lu, Junwen (2023). "Graph Neural Network-Based Efficient Subgraph Embedding Method for Link Prediction in Mobile Edge Computing". In: *Sensors*. DOI: 10.3390/s23104936.

Dimitriou, Paraskevas and Karyotis, Vasileios (2023). "A Combinatory Framework for Link Prediction in Complex Networks". In: *Applied Sciences*. DOI: 10.3390/app13179685.

Dimitriou, Paraskevas and Karyotis, Vasileios (2024). "Empowering Random Walk Link Prediction Algorithms in Complex Networks by Adapted Structural Information". In: *Ieee Access*. DOI: 10.1109/access.2024.3381510.

Djauhari, Maman A. and Gan, Siew Lee (2016). "Network Topology of Economic Sectors". In: *Journal of Statistical Mechanics Theory and Experiment*. DOI: 10.1088/1742-5468/2016/09/093401.

Domenico, Manlio De, Solé-Ribalta, Albert, Cozzo, Emanuele, Kivelä, Mikko, Moreno, Yamir, Porter, Mason A., Gómez, Sergio, and Arenas, Àlex (2013). "Mathematical Formulation of Multilayer Networks". In: *Physical Review X*. DOI: 10.1103/physrevx.3.041022.

Dong, Liyan, Li, Yongli, Yin, Han, Huang, Le, and Mao, Rui (2013). "The Algorithm of Link Prediction on Social Network". In: *Mathematical Problems in Engineering*. DOI: 10.1155/2013/125123.

Dueñas, Marco, Mastrandrea, Rossana, Barigozzi, Matteo, and Fagiolo, Giorgio (2017). "Spatio-temporal patterns of the international merger and acquisition network". In: *Scientific Reports* 7, p. 10789.

D'Agata, Rosario, Gozzo, Simona, and Tomaselli, Venera (2013). "Network analysis approach to map tourism mobility". In: *Quality & quantity* 47.6, pp. 3167–3184.

Enkel, Ellen, Gassmann, Oliver, and Chesbrough, Henry (2009). "Open R&D and open innovation: exploring the phenomenon". In: *R&d Management* 39.4, pp. 311–316. DOI: https://doi.org/10.1111/j.1467-9310.2009.00570.x.

Erdős, Paul and Rényi, Alfréd (1959). "On random graphs". In: *Publicationes Mathematicae* 6, pp. 290–297.

Erdős, P. and Rényi, A. (1959). "On random graphs, I". In: *Publicationes Mathematicae (Debrecen)* 6, pp. 290–297.

Erdős, P. and Rényi, A. (1960). "On the evolution of random graphs". In: *Publ. Math. Inst. Hung. Acad. Sci.* 5, pp. 17–61.

Erdős, P. and Rényi, A. (1961a). "On the evolution of random graphs". In: *Bull. Inst. Internat. Statist.* 38.4, pp. 343–347.

Erdős, P. and Rényi, A. (1961b). "On the Strength of Connectedness of a Random Graph". In: *Acta Mathematica Academiae Scientiarum Hungarica* 12, pp. 261–267.

Erdős, P. and Rényi, A. (1963). "Asymmetric graphs". In: *Acta Mathematica Academiae Scientiarum Hungarica* 14, pp. 295–315.

Erdős, P. and Rényi, A. (1966a). "On random matrices". In: *Publ. Math. Inst. Hung. Acad. Sci.* 8, pp. 455–461.

Erdős, P. and Rényi, A. (1966b). "On the existence of a factor of degree one of a connected random graph". In: *Acta Mathematica Academiae Scientiarum Hungarica* 17, pp. 359–368.

Erdős, P. and Rényi, A. (1968). "On random matrices II". In: *Studia Scientiarum Mathematicarum Hungarica* 3, pp. 459–464.

Expert, P., Evans, T. S., Blondel, V. D., and Lambiotte, R. (2011). "Uncovering space-independent communities in spatial networks". In: *Proceedings of the National Academy of Sciences* 108.19, 7663—7668. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.1018962108.

Fagiolo, Giorgio (2010). "The International-Trade Network: Gravity Equations and Topological Properties". In: *Journal of Economic Interaction and Coordination*. DOI: 10.1007/s11403-010-0061-y.

Fagiolo, Giorgio, Reyes, Javier, and Schiavo, Stefano (2010). "The evolution of the world trade web: a weighted-network analysis". In: *Journal of Evolutionary Economics* 20.4, pp. 479–514.

Farzaneh, Amirmohammad and Coon, Justin P. (2022). "An Information Theory Approach to Network Evolution Models". In: *arXiv preprint arXiv:2201.08306*. DOI: 10.48550/arxiv.2201.08306.

Fałdowski, Marek and Nepelski, Mariusz (2018). "EU Funds for Security". In: *Internal Security*. DOI: 10.5604/01.3001.0012.7497.

Fosdick, B. K., Larremore, D. B., Nishimura, J., and Ugander, J. (2018). "Configuring random graph models with fixed degree sequences". In: *SIAM Review* 60 (2), pp. 315–355. DOI: 10.1137/16m1087175.

Foster, J. G., Foster, D., Grassberger, P., and Paczuski, M. (2010). "Edge direction and the structure of networks". In: *Proceedings of the National Academy of Sciences* 107 (24), pp. 10815–10820. DOI: 10.1073/pnas.0912671107.

Frenken, Koen and Boschma, R.A. (2007). "A Theoretical Framework for Evolutionary Economic Geography: Industrial Dynamics and Urban Growth as a Branching Process". In: *Journal of Economic Geography*. DOI: 10.1093/jeg/lbm018.

Gadár, Laszló, Kosztyán, Zsolt T, and Abonyi, János (2018). "The Settlement Structure Is Reflected in Personal Investments: Distance-Dependent Network Modularity-Based Measurement of Regional Attractiveness". In: *Complexity* 2018, pp. 1–17. DOI: 10.1155/2018/1306704. URL: https://doi.org/10.1155/2018/1306704.

Gadar, Laszlo, Kosztyan, Zsolt T, and Abonyi, Janos (2018). "The settlement structure is reflected in personal investments: distance-dependent network modularity-based measurement of regional attractiveness". In: *Complexity* 2018, pp. 1–16. DOI: https://doi.org/10.1155/2018/1306704.

Gan, Jing, Zhang, Dongxue, Guo, Fuyou, and Dong, Erwei (2024). "Intensity of Tourism Economic Linkages in Chinese Land Border Cities and Network Characterization". In: *Sustainability*. DOI: 10.3390/su16051843.

Garcia-Bernardo, Javier, Fichtner, Jan, Takes, Frank W., and Heemskerk, Eelke M. (2017). "Uncovering Offshore Financial Centers: Conduits and Sinks in the Global Corporate Ownership Network". In: *Scientific Reports*. DOI: 10.1038/s41598-017-06322-9.

Ge, Xiou, Wang, Yun Cheng, Wang, Bin, Kuo, C-C Jay, et al. (2024). "Knowledge Graph Embedding: An Overview". In: *APSIPA Transactions on Signal and Information Processing* 13.1. DOI: https://doi.org/10.1561/116.00000065.

Georgiadou, Maria Christina (2018). "Renewable Fuels: The European Union Research and Innovation Policies". In: *Ecs Meeting Abstracts*. DOI: 10.1149/ma2018-02/54/2181.

Girvan, Michelle and Newman, Mark EJ (2002). "Community structure in social and biological networks". In: *Proceedings of the National Academy of Sciences* 99.12, pp. 7821–7826. DOI: https://doi.org/10.1073/pnas.122653799.

Glos, Adam, Krawiec, Aleksandra, and Pawela, Łukasz (2021). "Asymptotic Entropy of the Gibbs State of Complex Networks". In: *Scientific Reports*. DOI: 10.1038/s41598-020-78626-2.

Gold, Carl and Sollich, Peter (2003). "Model selection for support vector machine classification". In: *Neurocomputing* 55.1-2, pp. 221–249. DOI: https://doi.org/10.1016/S0925-2312(03)00375-8.

Goldstein, Bruce Evan and Butler, W. H. (2010). "Expanding the Scope and Impact of Collaborative Planning". In: *Journal of the American Planning Association*. DOI: 10.1080/01944361003646463.

Granovetter, Mark (2005). "The Impact of Social Structure on Economic Outcomes". In: *Journal of Economic Perspectives*. DOI: 10.1257/0895330053147958.

Gu, Quanquan, Li, Zhenhui, and Han, Jiawei (2011). "Linear discriminant dimensionality reduction". In: *Machine Learning and Knowledge Discovery in Databases:*

European Conference, ECML PKDD 2011, Athens, Greece, September 5-9, 2011. Proceedings, Part I 11. Springer, pp. 549–564. DOI: https://doi.org/10.1007/978-3-642-23780-5_45.

Gu, Weiwei, Hou, Jinqiang, and Gu, Weiyi (2023). "Improving Link Prediction Accuracy of Network Embedding Algorithms via Rich Node Attribute Information". In: *Journal of Social Computing*. DOI: 10.23919/jsc.2023.0018.

Guan, JianCheng, Zuo, KaiRui, Chen, KaiHua, and Yam, Richard CM (2016a). "Does country-level R&D efficiency benefit from the collaboration network structure?" In: *Research Policy* 45.4, pp. 770–784. DOI: https://doi.org/10.1016/j.respol.2016.01.003.

Guan, Qing, An, Haizhong, Gao, Xiangyun, Huang, Shupei, and Li, Huajiao (2016b). "Estimating potential trade links in the international crude oil trade: A link prediction approach". In: *Energy* 102, pp. 406–415. DOI: https://doi.org/10.1016/j.energy.2016.02.099.

Guns, Raf and Rousseau, Ronald (2014). "Recommending research collaborations using link prediction and random forest classifiers". In: *Scientometrics* 101, pp. 1461–1473. DOI: https://doi.org/10.1007/s11192-013-1228-9.

Guns, Raf and Wang, Lili (2017). "Detecting the emergence of new scientific collaboration links in Africa: A comparison of expected and realized collaboration intensities". In: *Journal of Informetrics* 11.3, pp. 892–903. DOI: https://doi.org/10.1016/j.joi.2017.07.004.

Hammoud, Zaynab and Krämer, Frank (2020). "Multilayer Networks: Aspects, Implementations, and Application in Biomedicine". In: *Big Data Analytics*. DOI: 10.1186/s41044-020-00046-0.

Han, Pu, Shi, Jin, Li, Xiaoyan, Wang, Dongbo, Shen, Si, and Su, Xinning (2014). "International collaboration in LIS: global trends and networks at the country and institution level". In: *Scientometrics* 98, pp. 53–72. DOI: https://doi.org/10.1007/s11192-013-1146-x.

Handcock, Mark S (2003). "Assessing degeneracy in statistical models of social networks". In: *Journal of the American Statistical Association* 76, pp. 33–50.

Hanneke, Steve, Fu, Wenjie, and Xing, Eric P (2010). "Discrete temporal models of social networks". In: *Electronic Journal of Statistics* 4, pp. 585–605.

Hasan, Mohammad Al and Zaki, Mohammed J (2011). "A survey of link prediction in social networks". In: *Social network data analytics*, pp. 243–275. DOI: https://doi.org/10.1007/978-1-4419-8462-3_9.

He, Xie, Ghasemian, Amir, Lee, Eun, Schwarze, Alice C., Clauset, Aaron, and Mucha, Peter J. (2024). "Link Prediction Accuracy on Real-World Networks Under Non-Uniform Missing-Edge Patterns". In: *Plos One*. DOI: 10.1371/journal.pone.0306883.

Head, Keith and Mayer, Thierry (2014). "Gravity equations: Workhorse, toolkit, and cookbook". In: *Handbook of International Economics* 4, pp. 131–195.

Hearst, Marti A., Dumais, Susan T, Osuna, Edgar, Platt, John, and Scholkopf, Bernhard (1998). "Support vector machines". In: *IEEE Intelligent Systems and their applications* 13.4, pp. 18–28. DOI: https://doi.org/10.1109/5254.708428.

Hogan, Stéphane (2017). "Horizon 2020: Opportunities for Medical Research and Innovation". In: *Annals of the Rheumatic Diseases*. DOI: 10.1136/annrheumdis-2017-eular.7225.

Huang, Liang, Li, Ruixuan, and Chen, Hong (2016). "Truncated Kernel Projection Machine for Link Prediction". In: *Journal of Computing Science and Engineering*. DOI: 10.5626/jcse.2016.10.2.58.

Huang, Lu, Zhu, Yihe, Zhang, Yi, Zhou, Xiao, and Jia, Xiang (2018). "A link prediction-based method for identifying potential cooperation partners: A case study on four journals of informetrics". In: *2018 Portland international conference on management of engineering and technology (PICMET)*. IEEE, pp. 1–6. DOI: https://doi.org/10.23919/PICMET.2018.8481974.

Hui, Eddie CM, Li, Xun, Chen, Tingting, and Lang, Wei (2020). "Deciphering the spatial structure of China's megacity region: A new bay area—The Guangdong-Hong Kong-Macao Greater Bay Area in the making". In: *Cities* 105, p. 102168.

Hunter, David R, Goodreau, Steven M, and Handcock, Mark S (2008). "Goodness of fit of social network models". In: *Journal of the American Statistical Association* 103.481, pp. 248–258.

Iwayama, Koji, Hirata, Yoshito, Takahashi, Kohske, Watanabe, Katsumi, Aihara, Kazuyuki, and Suzuki, Hideyuki (2012). "Characterizing Global Evolutions of Complex Systems via Intermediate Network Representations". In: *Scientific Reports*. DOI: 10.1038/srep00423.

Jackson, Matthew O (2008). *Social and economic networks*. Princeton University Press.

Jiang, Hong, Gao, Sipeng, Song, Yang, Sheng, Kuang, and Amaratunga, G.A.J. (2019). "An Empirical Study on the Impact of Collaborative R&D Networks on Enterprise Innovation Performance Based on the Mediating Effect of Technology Standard Setting". In: *Sustainability*. DOI: 10.3390/su11247249.

Jiang, Xin and Liang, Quanyi (2024). "Epidemic Process on Partially Overlapped Multi-Layer Networks". In: *Journal of Statistical Mechanics Theory and Experiment*. DOI: 10.1088/1742-5468/ad2dd7.

Johnston, Ron, Jones, Kelvyn, and Manley, David (2018). "Confounding and collinearity in regression analysis: a cautionary tale and an alternative procedure, illustrated by studies of British voting behaviour". In: *Quality & quantity* 52.4, pp. 1957–1976. DOI: https://doi.org/10.1007/s11135-017-0584-6.

Joy, Varghese, Shukla, A, and Jain, Vijay Kumar (2024). "Corporate Sustainability and Its Relevance for Business: A Bibliometric Approach". In: *Journal of Social Commerce*. DOI: 10.56209/jommerce.v4i1.84.

Jørgensen, Jacob Høj, Bergenholtz, Carsten, Goduscheit, René Chester, and Rasmussen, Erik Stavnsager (2011). "Managing Inter-Firm Collaboration in the Fuzzy Front-End: Structure as a Two-Edged Sword". In: *International Journal of Innovation Management*. DOI: 10.1142/s1363919611003118.

Kallioinen, Mika (June 2020). *Long-Distance Trade in Medieval Europe*. DOI: 10.1093/acrefore/9780190625979.013.558. URL: https://oxfordre.com/economics/view/10.1093/acrefore/9780190625979.001.0001/acrefore-9780190625979-e-558.

Katerndahl, David A. (2011). "Evolution of the Research Collaboration Network in a Productive Department". In: *Journal of Evaluation in Clinical Practice*. DOI: 10.1111/j.1365-2753.2011.01791.x.

Kim, Jounghyeon (2019). "Ownership Concentration and Institutional Quality: Do They Affect Corporate Bankruptcy Risk?" In: *Asia-Pacific Journal of Financial Studies*. DOI: 10.1111/ajfs.12271.

Kleinberg, Jon M. (Sept. 1999). "Authoritative sources in a hyperlinked environment". In: *J. ACM* 46.5, 604–632. ISSN: 0004-5411. DOI: 10.1145/324133.324140. URL: https://doi.org/10.1145/324133.324140.

Kong, Xiangjun, Wan, Jian-Bo, Hu, Hao, Su, Shi-Bing, and Y, Hu (2017). "Evolving Patterns in a Collaboration Network of Global R&D on Monoclonal Antibodies". In: *Mabs*. DOI: 10.1080/19420862.2017.1356527.

Kosztyán, Zsolt T, Katona, Attila I, Kuppens, Kurt, Kisgyörgy-Pál, Mária, Nachbagauer, Andreas, and Csizmadia, Tibor (2022a). "Exploring the structures and design effects of EU-funded R&D&I project portfolios". In: *Technological Forecasting and Social Change* 180, p. 121687. DOI: https://doi.org/10.1016/j.techfore.2022.121687.

Kosztyán, Zsolt Tibor, Csányi, Vivien Valéria, Banász, Zsuzsanna, Jakobi, Ákos, Neumanné-Virág, Ildikó, and Telcs, András (2021). "The role of higher education in spatial mobility". In: *Applied Network Science* 6.1, pp. 1–30. DOI: https://doi.org/10.1007/s41109-021-00428-w.

Kosztyán, Zsolt Tibor, Király, Ferenc, and Kurbucz, Marcell T (2022b). "Analysis of ownership network of European companies using gravity models". In: *Applied Network Science* 7.1, pp. 1–31. DOI: https://doi.org/10.1007/s41109-022-00501-y.

Kosztyán, Zsolt T., Király, Ferenc, Katona, Attila I., Csizmadia, Tibor, and Fehérvölgyi, Beáta (2024). "Analysis and prediction of the Horizon 2020 R&D&I collaboration network". In: *Expert Systems with Applications* 255, p. 124417. ISSN: 0957-4174. DOI: https://doi.org/10.1016/j.eswa.2024.124417. URL: https://www.sciencedirect.com/science/article/pii/S0957417424012831.

Kumar, Amit and Operti, Elisa (2023). "Missed Chances and Unfulfilled Hopes: Why Do Firms Make Errors in Evaluating Technological Opportunities?" In: *Strategic Management Journal*. DOI: 10.1002/smj.3543.

Kursa, Miron B, Jankowski, Aleksander, and Rudnicki, Witold R (2010). "Boruta–a system for feature selection". In: *Fundamenta Informaticae* 101.4, pp. 271–285. DOI: https://doi.org/10.3233/FI-2010-288.

Kurt, Yusuf and Kurt, Mustafa (2020). "Social network analysis in international business research: An assessment of the current state of play and future research directions". In: *International Business Review* 29.2, p. 101633.

Lacasa, Lucas, Nicosia, Vincenzo, and Latora, Vito (2015). "Network Structure of Multivariate Time Series". In: *Scientific Reports*. DOI: 10.1038/srep15508.

Lambert, Lisa Schurer and Newman, Daniel A. (2022). "Construct Development and Validation in Three Practical Steps: Recommendations for Reviewers, Editors, and Authors". In: *Organizational Research Methods*. DOI: 10.1177/10944281221115374.

Lande, Dmytro, Fu, Minglei, Guo, Wen, Balagura, Iryna, Gorbov, Ivan, and Yang, Hongbo (2020). "Link prediction of scientific collaboration networks based on information retrieval". In: *World Wide Web* 23, pp. 2239–2257. DOI: https://doi.org/10.1007/s11280-019-00768-9.

Laufs, Daniel, Melnychuk, Tetyana, and Schultz, Carsten (2024). "Effects of Prior Knowledge and Collaborations on R&D Performance in Times of Urgency: The Case of <scp>COVID</Scp>-19 Vaccine Development". In: *R and D Management*. DOI: 10.1111/radm.12670.

Lee, Duk Hee, Seo, Il Won, Choe, Ho Chull, and Kim, Hee Dae (2012). "Collaboration network patterns and research performance: the case of Korean public research institutions". In: *Scientometrics* 91.3, pp. 925–942. DOI: https://doi.org/10.1007/s11192-011-0602-8.

Li, Huan, Lu, Gang, and Guo, Junxia (Mar. 2015). "Generating Null Models for Large-Scale Networks on GPU". In: *Proceedings of the 2015 International Industrial Informatics and Computer Engineering Conference*. Atlantis Press, pp. 204–208. DOI: 10.2991/iiicec-15.2015.49. URL: https://doi.org/10.2991/iiicec-15.2015.49.

Li, Jie, Peng, Xiyang, Wang, Jian, and Zhao, Na (2021). "A Method for Improving the Accuracy of Link Prediction Algorithms". In: *Complexity*. DOI: 10.1155/2021/8889441.

Li, Taisong, Wang, Bing, Jiang, Yasong, Zhang, Yan, and Yan, Yonghong (2018). "Restricted Boltzmann machine-based approaches for link prediction in dynamic networks". In: *IEEE Access* 6, pp. 29940–29951. DOI: https://doi.org/10.1109/ACCESS.2018.2840054.

Li, Weihua, Zhang, Sam, Zheng, Zhiming, Cranmer, Skyler, and Clauset, Aaron (2022). "Untangling the Network Effects of Productivity and Prominence Among Scientists". In: *Nature Communications*. DOI: 10.1038/s41467-022-32604-6.

Li, Zhijun, Gao, Haiping, Shang, Zhiyong, and Zhang, Wenming (2023). "Robustness of Consensus of Two-Layer Ring Networks". In: *Symmetry*. DOI: 10.3390/sym15051085.

Liang, Huade, Zeng, Huilin, and Xiao-juan, Dong (2024). "Regional Economic Forecast Using Elman Neural Networks With Wavelet Function". In: *Plos One*. DOI: 10.1371/journal.pone.0299657.

Lidth Jeude, Jeroen van, Aste, Tomaso, and Caldarelli, Guido (2019). "The Multilayer Structure of Corporate Networks". In: *New Journal of Physics*. DOI: 10.1088/1367-2630/ab022d.

Liu, Fajian, Zhang, Jinhe, Zhang, Jie, Chen, Dongdong, Liu, Zehua, and Lu, Song (2012a). "Roles and functions of tourism destinations in tourism region of South Anhui: A tourist flow network perspective". In: *Chinese Geographical Science* 22.6, pp. 755–764.

Liu, Sen, Dong, Zhiliang, Ding, Chao, Wang, Tian, and Zhang, Yichi (2020). "Do you need cobalt ore? Estimating potential trade relations through link prediction". In: *Resources Policy* 66, p. 101632. DOI: https://doi.org/10.1016/j.resourpol.2020.101632.

Liu, Wei, Xiao, Haizhen, Xie, Renyi, Xi-ying, Luo, Yang, Xia, and Zhou, Jingyi (2024). "Does Government Ownership Help Make Private Firms Greener? Evidence From an Emerging Market". In: *Business Strategy and the Environment*. DOI: 10.1002/bse.3823.

Liu, Xin, Murata, Tsuyoshi, and Wakita, Ken (2012b). "Extending modularity by incorporating distance functions in the null model". In: *CoRR* abs/1210.4007, pp. 1–12.

Liu, Yangyang, Zhao, Chengli, Wang, Xiaojie, Huang, Qiangjuan, Zhang, Xue, and Yi, Dongyun (2016). "The degree-related clustering coefficient and its application to link prediction". In: *Physica A: Statistical Mechanics and Its Applications* 454, pp. 24–33. DOI: https://doi.org/10.1016/j.physa.2016.02.014.

Liu, Zeguang, Li, Yao, and Liu, Huilin (2019). "Link Prediction in Evolving Networks Base on Information Propagation". In: *Ieee Access*. DOI: 10.1109/access.2019.2942357.

Liu, Zhenyuan, Mu, Renyan, Hu, Shuhua, Li, Mengqi, and Wang, Li (2018). "The method and application of graphic recognition of the social network structure of urban agglomeration". In: *Wireless Personal Communications* 103.1, pp. 447–480.

Lobo, José, A. Bettencourt, Luís M., Strumsky, Deborah, and West, Geoffrey B. (2013). "Urban Scaling and the Production Function for Cities". In: *Plos One*. DOI: 10.1371/journal.pone.0058407.

Long, Janet C., Hibbert, Peter, and Braithwaite, Jeffrey (2015). "Structuring Successful Collaboration: A Longitudinal Social Network Analysis of a Translational Research Network". In: *Implementation Science*. DOI: 10.1186/s13012-016-0381-y.

Lusher, Dean, Koskinen, Johan, and Robins, Garry (2013). *Exponential random graph models for social networks: Theory, methods, and applications*. Cambridge University Press.

Magnusson, Thomas and Werner, Viktor (2022). "Conceptualisations of Incumbent Firms in Sustainability Transitions: Insights From Organisation Theory and a Systematic Literature Review". In: *Business Strategy and the Environment*. DOI: 10.1002/bse.3081.

Mahnken, Torben A and Moehrle, Martin G (2018). "Multi-cross-industry innovation patents in the USA-A combination of PATSTAT and Orbis search". In: *World Patent Information* 55, pp. 52–60. DOI: https://doi.org/10.1016/j.wpi.2018.10.003.

Mantovani, Rafael G, Rossi, André LD, Vanschoren, Joaquin, Bischl, Bernd, and De Carvalho, André CPLF (2015). "Effectiveness of random search in SVM hyper-parameter tuning". In: *2015 International Joint Conference on Neural Networks (IJCNN)*. Ieee, pp. 1–8. DOI: https://doi.org/10.1109/IJCNN.2015.7280664.

Mao, Min and Cheng, Xi (2019). "Evolution Analysis of Foreign Trade Network Structructure Based on Complex Network SNA". In: *Proceedings of the 2019 2nd International Conference on E-Business, Information Management and Computer Science*, pp. 1–5.

Mastrandrea, Rossana, Squartini, Tiziano, Fagiolo, Giorgio, and Garlaschelli, Diego (2014). "Enhanced reconstruction of weighted networks from strengths and degrees". In: *New Journal of Physics* 16.4, p. 043022.

McCarthy, Seán (2018). "Success Rates in Horizon 2020". In: *Journal of Innovation Management*. DOI: 10.24840/2183-0606_005.004_0003.

Meng, Xiangyi and Zhou, Bin (2023). "Scale-free networks beyond power-law degree distribution". In: *Chaos, Solitons & Fractals* 176, p. 114173. ISSN: 0960-0779. DOI: https://doi.org/10.1016/j.chaos.2023.114173. URL: https://www.sciencedirect.com/science/article/pii/S0960077923010755.

Međedović, Janko (2020). "Human Life Histories as Dynamic Networks: Using Network Analysis to Conceptualize and Analyze Life History Data". In: *Evolutionary Psychological Science*. DOI: 10.1007/s40806-020-00252-y.

Milgram, S. (1967). "The Small-World Problem". In: *Psychology Today* 1.1, pp. 61–67. DOI: 10.1037/e400002009-005.

Mizuno, Takayuki, Doi, Shohei, and Kurizaki, Shuhei (2020). "The Power of Corporate Control in the Global Ownership Network". In: *Plos One*. DOI: 10.1371/journal.pone.0237862.

Mizuno, Takayuki, Doi, Shohei, and Kurizaki, Shuhei (2023). "The Flow of Corporate Control in the Global Ownership Network". In: *Plos One*. DOI: 10.1371/journal.pone.0290229.

Mohan, Anuraj, Venkatesan, R, and Pramod, KV (2017). "A scalable method for link prediction in large real world networks". In: *Journal of Parallel and Distributed Computing* 109, pp. 89–101. DOI: https://doi.org/10.1016/j.jpdc.2017.05.009.

Montoro Sánchez, María Ángeles, Ortiz-de-Urbina-Criado, Marta, and Mora Valentín, Eva María (2011). "Effects of Knowledge Spillovers on Innovation and Collaboration in Science and Technology Parks". In: *Journal of Knowledge Management*. DOI: 10.1108/13673271111179307.

Morrison, Andrea (2008). "Gatekeepers of knowledge within industrial districts: who they are, how they interact". In: *Regional Studies* 42.6, pp. 817–835.

Mou, Naixia, Zheng, Yunhao, Makkonen, Teemu, Yang, Tengfei, Tang, Jinwen Jimmy, and Song, Yan (2020). "Tourists' digital footprint: The spatial patterns of tourist flows in Qingdao, China". In: *Tourism Management* 81, p. 104151.

Moutinho, João P., Magano, Duarte, and Coutinho, Bruno (2024). "On the Complexity of Quantum Link Prediction in Complex Networks". In: *Scientific Reports*. DOI: 10.1038/s41598-023-49906-4.

Nakamoto, Tembo, Chakraborty, Abhijit, and Ikeda, Yuichi (2019). "Identification of Key Companies for International Profit Shifting in the Global Ownership Network". In: *Applied Network Science*. DOI: 10.1007/s41109-019-0158-8.

Natekin, Alexey and Knoll, Alois (2013). "Gradient boosting machines, a tutorial". In: *Frontiers in neurorobotics* 7, p. 21. DOI: https://doi.org/10.3389/fnbot.2013.00021.

Nejad, Ahmad Jafar, Nejad, Jafar Bagheri, and Sepehri, Sepehr (2013). "Canonical Correlation Analysis Between Collaborative Networks and Innovation: A Case Study in Information Technology Companies in Province of Tehran, Iran". In: *Management Science Letters*. DOI: 10.5267/j.msl.2013.06.004.

Newman, M. E. J. (2003). "The Structure and Function of Complex Networks". In: *SIAM Review* 45.2, pp. 167–256. DOI: 10.1137/S003614450342480.

Newman, M. E. J. and Girvan, M. (2004). "Finding and evaluating community structure in networks". In: *Physical Review E* 69.2, p. 026113. DOI: 10.1103/PhysRevE.69.026113.

Newman, Mark (2010a). "Networks: an introduction". In: *Oxford University Press*.

Newman, Mark (2010b). *Networks: An Introduction*. Google-Books-ID: q7HVtpYVfC0C. OUP Oxford. ISBN: 978-0-19-920665-0.

Novitzky, Peter et al. (2020). "Improve alignment of research policy and societal values". In: *Science* 369.6499, pp. 39–41. DOI: https://doi.org/10.1007/s11192-013-1146-x/10.1126/science.abb3415.

Ntim, Collins G., Opong, Kwaku K., Danbolt, Jo, and Thomas, Dennis (2012). "Voluntary Corporate Governance Disclosures by Post-Apartheid South African Corporations". In: *Journal of Applied Accounting Research*. DOI: 10.1108/09675421211254830.

Ozcan, Sercan and Islam, Nazrul (2014). "Collaborative Networks and Technology Clusters — The Case of Nanowire". In: *Technological Forecasting and Social Change*. DOI: 10.1016/j.techfore.2013.08.008.

Ozer, Mine, Demirkan, Irem, and Gokalp, Omer N. (2013). "Collaboration Networks and Innovation: Does Corporate Lobbying Matter?" In: *Journal of Strategy and Management*. DOI: 10.1108/jsma-01-2013-0009.

Paas, Tiiu, Tafenau, Egle, and Scannell, Nancy J (2008). "Gravity equation analysis in the context of international trade: Model specification implications in the case of the European Union". In: *Eastern European Economics* 46.5, pp. 92–113.

Pal, Mahesh (2005). "Random forest classifier for remote sensing classification". In: *International journal of remote sensing* 26.1, pp. 217–222. DOI: https://doi.org/10.1080/01431160412331269698.

Parisi, Federica, Caldarelli, Guido, and Squartini, Tiziano (2018). "Entropy-Based Approach to Missing-Links Prediction". In: *Applied Network Science*. DOI: 10.1007/s41109-018-0073-4.

Pedarsani, Pedram and Grossglauser, Matthias (2011). "On the Privacy of Anonymized Networks". In: *ACM SIGMETRICS Performance Evaluation Review* 39.1, pp. 75–76. DOI: 10.1145/2020408.2020596.

Pehrsson, Anders (2016). "Sequential Expansion in a Foreign Market". In: *European Business Review*. DOI: 10.1108/ebr-01-2016-0017.

Pilosof, Shai, Porter, Mason A., Pascual, Mercedes, and Kéfi, Sonia (2017). "The Multilayer Nature of Ecological Networks". In: *Nature Ecology & Evolution*. DOI: 10.1038/s41559-017-0101.

Pinto Leão, Pedro Henrique and Silva, Miguel Mira da (2021). "Impacts of Digital Transformation on Firms' Competitive Advantages: A Systematic Literature Review". In: *Strategic Change*. DOI: 10.1002/jsc.2459.

Pişcoran, L. (2021). "Nonholonomic frame for a deformed (α, β)-metric". In: *International Electronic Journal of Geometry* 14 (2), pp. 231–238. DOI: 10.36890/iejg.973879.

Probst, Philipp, Wright, Marvin N, and Boulesteix, Anne-Laure (2019). "Hyperparameters and tuning strategies for random forest". In: *Wiley Interdisciplinary Reviews: data mining and knowledge discovery* 9.3, e1301. DOI: https://doi.org/10.1002/widm.1301.

Puślecki, Zdzisław W. (2016). "The Financial Instrument of the Innovation Union — Horizon 2020". In: *Rocznik Europeistyczny*. DOI: 10.19195/2450-274x.2.2.

Pérez, Leobardo Plata, Sánchez-Pérez, Joss, and Sánchez-Sánchez, Francisco (2015). "An Elementary Characterization of the Gini Index". In: *Mathematical Social Sciences*. DOI: 10.1016/j.mathsocsci.2015.01.002.

Qi, Yan, Zhang, Xin, Hu, Zhengyin, Xiang, Bin, Zhang, Ran, and Fang, Shu (2022). "Choosing the right collaboration partner for innovation: a framework based on topic analysis and link prediction". In: *Scientometrics* 127.9, pp. 5519–5550. DOI: https://doi.org/10.1007/s11192-022-04306-9.

Ren, Z., Pan, X., and Zhang, Y. (2020). "Significance of the nested structure in multiplex world trade networks". In: *Complexity* 2020, pp. 1–9. DOI: 10.1155/2020/8827840.

Reyes, Javier A., Wooster, Rossitza B., and Shirrell, Stuart (2014). "Regional Trade Agreements and the Pattern of Trade: A Networks Approach". In: *World Economy*. DOI: 10.1111/twec.12121.

Rigtering, Coen and Behrens, Martin (2021). "The Effect of Corporate — Start-Up Collaborations on Corporate Entrepreneurship". In: *Review of Managerial Science*. DOI: 10.1007/s11846-021-00443-2.

Rinaldo, Alessandro, Petrović, Sonja, and Fienberg, Stephen E. (2013). "Maximum Likelihood Estimation in the β-Model". In: *The Annals of Statistics* 41.5, pp. 2246–2272. DOI: 10.1214/12-aos1078.

Robins, Garry, Pattison, Pip, Kalish, Yuval, and Lusher, Dean (2007). "An introduction to exponential random graph (p*) models for social networks". In: *Social Networks* 29.2, pp. 173–191.

Roediger-Schluga, Thomas and Barber, Michael J (2008). "R&D collaboration networks in the European Framework Programmes: Data processing, network construction and selected results". In: *International Journal of Foresight and Innovation Policy* 4.3-4, pp. 321–347. DOI: https://doi.org/10.1504/IJFIP.2008.017583.

Roesler, Christoph and Broekel, Tom (2017). "The Role of Universities in A network of Subsidized R&D collaboration: The Case of the Biotechnology-Industry in Germany". In: *Review of Regional Research*. DOI: 10.1007/s10037-017-0118-7.

Romei, Andrea, Ruggieri, Salvatore, and Turini, Franco (2015). "The layered structure of company share networks". In: *2015 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, pp. 1–10. DOI: 10.1109/DSAA.2015.7344809.

Rossi, Andrea, Barbosa, Denilson, Firmani, Donatella, Matinata, Antonio, and Merialdo, Paolo (2021). "Knowledge graph embedding for link prediction: A comparative analysis". In: *ACM Transactions on Knowledge Discovery from Data (TKDD)* 15.2, pp. 1–49. DOI: https://doi.org/10.1145/3424672.

Rungi, Armando, Morrison, Gregory, and Pammolli, Fabio (2017). "Global Ownership and Corporate Control Networks". In: *SSRN Electronic Journal*. DOI: 10.2139/ssrn.3031955.

Sagi, Omer and Rokach, Lior (2018). "Ensemble learning: A survey". In: *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 8.4, e1249. DOI: https://doi.org/10.1002/widm.1249.

Saletti, Costanza, Morini, Mirko, and Gambarotta, Agostino (2020). "The Status of Research and Innovation on Heating and Cooling Networks as Smart Energy Systems Within Horizon 2020". In: *Energies*. DOI: 10.3390/en13112835.

Santos Silva, JMC and Tenreyro, Silvana (2006). "The log of gravity". In: *The Review of Economics and Statistics* 88.4, pp. 641–658.

Scherngell, Thomas and Lata, Rafael (2013). "Towards an integrated European Research Area? Findings from Eigenvector spatially filtered spatial interaction models using European Framework Programme data". In: *Papers in Regional Science* 92.3, pp. 555–577. DOI: https://doi.org/10.1111/j.1435-5957.2012.00419.x.

Schweitzer, Frank, Fagiolo, Giorgio, Sornette, Didier, Vega-Redondo, Fernando, Vespignani, Alessandro, and White, Douglas R (2009). "Economic networks: The new challenges". In: *Science* 325.5939, pp. 422–425.

Searle, Glen, Sigler, Thomas, and Martinus, Kirsten (2018). "Firm evolution and cluster specialization: a social network analysis of resource industry change in two Australian cities". In: *Regional studies, regional science* 5.1, pp. 369–387.

Sebestyén, Tamás and Varga, Attila (2013). "Research productivity and the quality of interregional knowledge networks". In: *The Annals of Regional Science* 51.1, pp. 155–189.

Secundo, Giustina, Toma, Antonio, Schiuma, Giovanni, and Passiante, Giuseppina (2019). "Knowledge transfer in open innovation: A classification framework for healthcare ecosystems". In: *Business Process Management Journal* 25.1, pp. 144–163. DOI: https://doi.org/10.1108/BPMJ-06-2017-0173.

Seok, Hwayoon, Barnett, George A, and Nam, Yoonjae (2021). "A social network analysis of international tourism flow". In: *Quality & Quantity* 55.2, pp. 419–439.

Serrano, M Angeles and Boguñá, Marián (2003). "Topology of the world trade web". In: *Physical Review E* 68.1, p. 015101.

Shadish, William R. (2010). "Campbell and Rubin: A Primer and Comparison of Their Approaches to Causal Inference in Field Settings." In: *Psychological Methods*. DOI: 10.1037/a0015916.

Shahriari, Bobak, Swersky, Kevin, Wang, Ziyu, Adams, Ryan P, and De Freitas, Nando (2015). "Taking the human out of the loop: A review of Bayesian optimization". In: *Proceedings of the IEEE* 104.1, pp. 148–175. DOI: https://doi.org/10.1109/JPROC.2015.2494218.

Shao, Jingbo, Liu, Xiaoxiao, Li, Yingmei, and Liu, Jingyu (2015). "Database Performance Optimization for SQL Server Based on Hierarchical Queuing Network Model". In: *International Journal of Database Theory and Application*. DOI: 10.14257/ijdta.2015.8.1.19.

Sharifi, Farnoush Amini and Jafari, Seyedeh Mahbubeh (2016). "Cash Flows and Leverage Adjustments". In: *Accounting*. DOI: 10.5267/j.ac.2016.4.001.

Sharma, Kiran, Chakrabarti, Anindya S., and Chakraborti, Anirban (2019). "Multi-Layered Network Structure: Relationship Between Financial and Macroeconomic Dynamics". In: *New Economic Windows*. DOI: 10.1007/978-3-030-11364-3_9.

Sharp, Lucy (2019). "Playing a Critical Role in Expanding Europe's Horizons". In: *Impact*. DOI: 10.21820/23987073.2019.8.4.

Sheridan, Robert P, Wang, Wei Min, Liaw, Andy, Ma, Junshui, and Gifford, Eric M (2016). "Extreme gradient boosting as a method for quantitative structure–activity relationships". In: *Journal of chemical information and modeling* 56.12, pp. 2353–2360. DOI: https://doi.org/10.1021/acs.jcim.6b00591.

Shi, Jianbang and Xiao, Zhenhong (2024). "Research on the Impact of Inter-Industry Innovation Networks on Collaborative Innovation Performance: A Case Study of Strategic Emerging Industries". In: *Systems*. DOI: 10.3390/systems12060211.

Sigler, Thomas and Martinus, Kirsten (2016). "Extending Beyond 'World Cities' in World City Network (WCN) Research: Urban Positionality and Economic Linkages Through the Australia-Based Corporate Network". In: *Environment and Planning a Economy and Space*. DOI: 10.1177/0308518x16659478.

Singh, Harmeet (2023). "A Novel Hybrid Approach for Similarity-Based Link Prediction in Complex Networks". In: *TJJPT*. DOI: 10.52783/tjjpt.v44.i4.1134.

Sjøberg, Dag I.K. and Bergersen, Gunnar R. (2023). "Improving the Reporting of Threats to Construct Validity". In: *Proceedings of the 27th International Conference on Evaluation and Assessment in Software Engineering*. EASE '23. Oulu, Finland: Association for Computing Machinery, 205–209. ISBN: 9798400700446. DOI: 10.1145/3593434.3593449. URL: https://doi.org/10.1145/3593434.3593449.

Snijders, Tom AB (2011). "Statistical models for social networks". In: *Annual Review of Sociology* 37, pp. 131–153.

Snijders, Tom AB, Pattison, Philippa E, Robins, Garry L, and Handcock, Mark S (2006). "New specifications for exponential random graph models". In: *Sociological Methodology* 36.1, pp. 99–153.

Song, Xizhuoran, Zhang, Yan, Pan, Rui, and Wang, Hansheng (2022). *Link Prediction for Statistical Collaboration Networks Incorporating Institutes and Research Interests*. DOI: 10.1109/access.2022.3210129.

Spender, John-Christopher, Corvello, Vincenzo, Grimaldi, Michele, and Rippa, Pierluigi (2017). "Startups and Open Innovation: A Review of the Literature". In: *European Journal of Innovation Management*. DOI: 10.1108/ejim-12-2015-0131.

Squartini, Tiziano and Garlaschelli, Diego (2011). "Analytical maximum-likelihood method to detect patterns in real networks". In: *New Journal of Physics* 13.8, p. 083001.

Squartini, Tiziano and Garlaschelli, Diego (2018). "Maximum-entropy networks: Pattern detection, network reconstruction and graph combinatorics". In: *Springer*.

Squartini, Tiziano, Mol, Joey de, Hollander, Frank den, and Garlaschelli, Diego (2015). "Breaking of ensemble equivalence in networks". In: *Physical Review Letters* 115.26, p. 268701.

Steiber, Annika (2020). "Technology Management: Corporate-Startup Co-Location and How to Measure the Effects". In: *Journal of Technology Management & Innovation*. DOI: 10.4067/s0718-27242020000200011.

Steiber, Annika and Alänge, Sverker (2020). "Corporate-Startup Collaboration: Effects on Large Firms' Business Transformation". In: *European Journal of Innovation Management*. DOI: 10.1108/ejim-10-2019-0312.

Sulaimany, Sadegh, Khansari, Mohammad, Zarrineh, Peyman, Daianu, Madelaine, Jahanshad, Neda, Thompson, Paul M., and Masoudi-Nejad, Ali (2017). "Predicting Brain Network Changes in Alzheimer's Disease With Link Prediction Algorithms". In: *Molecular Biosystems*. DOI: 10.1039/c6mb00815a.

Sun, Qingshuang, Hu, Rongjing, Yang, Zhao, Yao, Yabing, and Yang, Fan (2017). "An improved link prediction algorithm based on degrees and similarities of nodes". In: *2017 IEEE/ACIS 16th International Conference on Computer and Information Science (ICIS)*. IEEE, pp. 13–18. DOI: https://doi.org/10.1109/ICIS.2017.7959962.

Sun, Shixiang, Wei, Xinjiang, Zhang, Huifeng, and Hu, Xin (2024). "Finite-time Pinning Synchronization Control for Multi-layer Complex Networks". In: *International Journal of Robust and Nonlinear Control*. DOI: 10.1002/rnc.7453.

Takes, Frank W., Kosters, Walter A., Witte, Boyd, and Heemskerk, Eelke M. (2018). "Multiplex Network Motifs as Building Blocks of Corporate Networks". In: *Applied Network Science*. DOI: 10.1007/s41109-018-0094-z.

Tang, Bo, Chen, Zehui, Zhang, Yuanyuan, and Hua, Sun (2022). "A Study on the Evolution of Economic Patterns and Urban Network System in Guangdong-Hong

Kong-Macao Greater Bay Area". In: *Frontiers in Public Health*. DOI: 10.3389/fpubh.2022.973843.

Tarasconi, Gianluca and Menon, Carlo (2017). *Matching Crunchbase with patent data*. Tech. rep. OECD. DOI: https://doi.org/10.1787/15f967fa-en.

Tharwat, Alaa (2016). "Linear vs. quadratic discriminant analysis classifier: a tutorial". In: *International Journal of Applied Pattern Recognition* 3.2, pp. 145–180. DOI: https://doi.org/10.1504/IJAPR.2016.079050.

Tian, Mingyu, Su, Yiwei, and Yang, Zhong (2021). "University–industry Collaboration and Firm Innovation: An Empirical Study of the Biopharmaceutical Industry". In: *The Journal of Technology Transfer*. DOI: 10.1007/s10961-021-09877-y.

Tonkin, Matthew, Woodhams, Jessica, Bull, Ray, Bond, John W, and Santtila, Pekka (2012). "A comparison of logistic regression and classification tree analysis for behavioural case linkage". In: *Journal of Investigative Psychology and Offender Profiling* 9.3, pp. 235–258. DOI: https://doi.org/10.1002/jip.1367.

Traag, V. A., Waltman, L., and Eck, N. J. van (Mar. 26, 2019). "From Louvain to Leiden: guaranteeing well-connected communities". In: *Scientific Reports* 9.1, p. 5233. ISSN: 2045-2322. DOI: 10.1038/s41598-019-41695-z. URL: https://doi.org/10.1038/s41598-019-41695-z.

Turner, Ryan, Eriksson, David, McCourt, Michael, Kiili, Juha, Laaksonen, Eero, Xu, Zhen, and Guyon, Isabelle (2021). "Bayesian Optimization is Superior to Random Search for Machine Learning Hyperparameter Tuning: Analysis of the Black-Box Optimization Challenge 2020". In: *Proceedings of the NeurIPS 2020 Competition and Demonstration Track*. Ed. by Hugo Jair Escalante and Katja Hofmann. Vol. 133. Proceedings of Machine Learning Research. PMLR, pp. 3–26. URL: https://proceedings.mlr.press/v133/turner21a.html.

Uhrig, Bettina (2019). "Impact of Social Sciences and Humanities for a European Research Agenda - Valuation of SSH in Mission Oriented Research; Rethinking Societal Impact – Collaboration With Stakeholders". In: DOI: 10.22163/fteval.2019.378.

Uukkivi, Raigo and Koppel, Ott (2020). "Assessment of the Economic Regulation of Network Industries: Oil Shale Value Chain in Estonia". In: *Oil Shale*. DOI: 10.3176/oil.2020.2.05.

Uukkivi, Raigo, Ots, Märt, and Koppel, Ott (2014). "Systematic Approach to Economic Regulation of Network Industries in Estonia". In: *Trames Journal of the Humanities and Social Sciences*. DOI: 10.3176/tr.2014.3.02.

Van Meeteren, Michiel, Neal, Zachary, and Derudder, Ben (2016). "Disentangling agglomeration and network externalities: A conceptual typology". In: *Papers in Regional Science* 95.1, pp. 61–80.

Vanni, Tázio, Mesa-Frias, Marco, Sánchez-García, Rubén J., Roesler, Rafael, Schwartsmann, Gilberto, Goldani, Marcelo Zubarán, and Foss, Anna M. (2014). "International Scientific Collaboration in HIV and HPV: A Network Analysis". In: *Plos One*. DOI: 10.1371/journal.pone.0093376.

Veugelers, Reinhilde, Cincera, Michele, Frietsch, Rainer, Rammer, Christian, Schubert, Torben, Pelle, Anita, Renda, Andrea, Montalvo, Carlos, and Leijten, Jos (2015). "The impact of horizon 2020 on innovation in Europe". In: *Intereconomics* 50.1, pp. 4–30. DOI: https://doi.org/10.1007/s10272-015-0521-7.

Vida, Bianka (2020). "Policy Framing and Resistance: Gender Mainstreaming in Horizon 2020". In: *European Journal of Women S Studies*. DOI: 10.1177/1350506820935495.

Villamil, Isabela, Kertész, János, and Fazekas, Mihály (2024). "Collusion Risk in Corporate Networks". In: *Scientific Reports*. DOI: 10.1038/s41598-024-53625-9.

Vitali, Stefania and Battiston, Stefano (2013). "The Community Structure of the Global Corporate Network". In: *SSRN Electronic Journal*. DOI: 10.2139/ssrn.2198974.

Vitali, Stefania and Battiston, Stefano (2014). "The Community Structure of the Global Corporate Network". In: *Plos One*. DOI: 10.1371/journal.pone.0104655.

Vitali, Stefania, Glattfelder, James B., and Battiston, Stefano (2011). "The Network of Global Corporate Control". In: *Plos One*. DOI: 10.1371/journal.pone.0025995.

Voitalov, Ivan, Hoorn, Pim van der, Hofstad, Remco van der, and Krioukov, Dmitri (2019). "Scale-free networks well done". In: *Phys. Rev. Res.* 1 (3), p. 033034. DOI: 10.1103/PhysRevResearch.1.033034. URL: https://link.aps.org/doi/10.1103/PhysRevResearch.1.033034.

Wall, Ronald and Knaap, G. A. van (2011). "Sectoral Differentiation and Network Structure Within Contemporary Worldwide Corporate Networks". In: *Economic Geography*. DOI: 10.1111/j.1944-8287.2011.01122.x.

Walrand, Jean (2008). "Economic Models of Communication Networks". In: *Performance Modeling and Engineering*. Ed. by Zhen Liu and Cathy H. Xia. Boston, MA: Springer US, pp. 57–89. ISBN: 978-0-387-79361-0. DOI: 10.1007/978-0-387-79361-0_3. URL: https://doi.org/10.1007/978-0-387-79361-0_3.

Walton-Roberts, Margaret (2011). "Immigration, Trade and 'Ethnic Surplus Value': A Critique of Indo–Canadian Transnational Networks". In: *Global Networks*. DOI: 10.1111/j.1471-0374.2011.00318.x.

Wang, Jingwei, Ma, Yunlong, Liu, Min, and Shen, Weiming (2019a). "Link Prediction Based on Community Information and Its Parallelization". In: *Ieee Access*. DOI: 10.1109/access.2019.2907202.

Wang, Lei, Ren, Jing, Xu, Bo, Li, Jianxin, Luo, Wei, and Xia, Feng (2020). "MODEL: Motif-Based Deep Feature Learning for Link Prediction". In: *Ieee Transactions on Computational Social Systems*. DOI: 10.1109/tcss.2019.2962819.

Wang, Peng, Xu, BaoWen, Wu, YuRong, and Zhou, XiaoYu (2015a). "Link prediction in social networks: the state-of-the-art". In: *Science China Information Sciences* 58.1, pp. 1–38. ISSN: 1869-1919. DOI: 10.1007/s11432-014-5237-y. URL: https://doi.org/10.1007/s11432-014-5237-y.

Wang, Xingxing, Wang, Anjian, and Zhu, Depeng (2022). "Simulation Analysis of Supply Crisis Propagation Based on Global Nickel Industry Chain". In: *Frontiers in Energy Research*. DOI: 10.3389/fenrg.2022.919510.

Wang, ZB, Han, WM, Sun, ZM, and Pan, XL (2019b). "Research on scientific collaboration prediction based on the combination of network topology and node attributes". In: *Information Studies: Theory & Application* 42.8, pp. 116–120.

Wang, Zhen, Wang, Lin, Szolnoki, Attila, and Perc, Matjaž (2015b). "Evolutionary Games on Multilayer Networks: A Colloquium". In: *The European Physical Journal B*. DOI: 10.1140/epjb/e2015-60270-7.

Watts, Duncan J. and Strogatz, Steven H. (1998). "Collective dynamics of 'small-world' networks". In: *Nature* 393.6684, pp. 440–442. ISSN: 1476-4687. DOI: 10.1038/30918. URL: https://doi.org/10.1038/30918.

Weidenfeld, Adi, Makkonen, Teemu, and Clifton, Nick (2021). "From interregional knowledge networks to systems". In: *Technological Forecasting and Social Change* 171, p. 120904.

Wider, Nicolas, Garas, Antonios, Scholtes, Ingo, and Schweitzer, Frank (2016). "An Ensemble Perspective on Multi-Layer Networks". In: DOI: 10.1007/978-3-319-23947-7_3.

Wohlin, Claes, Runeson, Per, Höst, Martin, Ohlsson, Magnus C, Regnell, Björn, and Wesslén, Anders (2012). *Experimentation in software engineering*. Springer Science & Business Media.

Xie, Wujie, Hai-jian, LI, and, Yufang Yin (2021). "Research on the Spatial Structure of the European Union's Tourism Economy and Its Effects". In: *International Journal of Environmental Research and Public Health*. DOI: 10.3390/ijerph18041389.

Xin, Peiming and Zhao, Hong (2009). "Time series forecasting using multilayer neural network constructed by a Monte-Carlo based algorithm". In: *2009 1st IEEE Symposium on Web Society*, pp. 264–267. DOI: 10.1109/SWS.2009.5271810.

Xu, Min and Yin, Yongchao (2017). "A similarity index algorithm for link prediction". In: *2017 12th International Conference on Intelligent Systems and Knowledge Engineering (ISKE)*. IEEE, pp. 1–6. DOI: https://doi.org/10.1109/ISKE.2017.8258724.

Yang, Jaewon and Leskovec, Jure (2015). "Defining and evaluating network communities based on ground-truth". In: *Knowledge and Information Systems* 42.1, pp. 181–213.

Yang, Qian, Dong, Enming, and Xie, Zheng (2014). "Link prediction via nonnegative matrix factorization enhanced by blocks information". In: *2014 10th International Conference on Natural Computation (ICNC)*. IEEE, pp. 823–827. DOI: https://doi.org/10.1109/ICNC.2014.6975944.

Ye, Cong, Slavakis, Konstantinos, Nakuci, Johan, Muldoon, Sarah F., and Medaglia, John (2021). "Online Classification of Dynamic Multilayer-Network Time Series in Riemannian Manifolds". In: *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 3815–3819. DOI: 10.1109/ICASSP39728.2021.9413560.

Ye, Meng, Mao, Wenjie, and Wang, Yaqi (2022). *The Spatial Structure of Regional Logistics and Influencing Factors: An Empirical Analysis Based on Sichuan Province, China*. DOI: 10.21203/rs.3.rs-1213810/v1.

Yu, Dejian, Xu, Zeshui, and Wang, Xizhao (2020). "Bibliometric analysis of support vector machines research trend: a case study in China". In: *International Journal of Machine Learning and Cybernetics* 11, pp. 715–728. DOI: https://doi.org/10.1007/s13042-019-01028-y.

Yuliansyah, Herman, Othman, Zulaiha Ali, and Bakar, Azuraliza Abu (2020). "Taxonomy of Link Prediction for Social Network Analysis: A Review". In: *Ieee Access*. DOI: 10.1109/access.2020.3029122.

Zanin, Massimiliano (2015). "Can We Neglect the Multi-Layer Structure of Functional Networks?" In: *Physica a Statistical Mechanics and Its Applications*. DOI: 10.1016/j.physa.2015.02.099.

Zattoni, Alessandro (2011). "Who Should Control a Corporation? Toward a Contingency Stakeholder Model for Allocating Ownership Rights". In: *Journal of Business Ethics*. DOI: 10.1007/s10551-011-0864-3.

Zeng, Yujie, Liu, Bo, Zhou, Fang, and Lü, Linyuan (2023). "Hyper-Null Models and Their Applications". In: *Entropy*. DOI: 10.3390/e25101390.

Zhang, Chengjun, Li, Qi, Lei, Yi, Qian, Ming, Shen, Xinyu, Cheng, Di, and Yu, Wenbin (2023). "The Absence of a Weak-Tie Effect When Predicting Large-Weight Links in Complex Networks". In: *Entropy*. DOI: 10.3390/e25030422.

Zhang, Chuanting, Zhang, Haixia, Yuan, Dongfeng, and Zhang, Minggao (2016a). "Deep learning based link prediction with social pattern and external attribute knowledge in bibliographic networks". In: *2016 IEEE International Conference on Internet of Things (iThings) and IEEE Green Computing and Communications (GreenCom) and IEEE Cyber, Physical and Social Computing (CPSCom) and IEEE Smart Data (SmartData)*. IEEE, pp. 815–821. DOI: https://doi.org/10.1109/iThings-GreenCom-CPSCom-SmartData.2016.170.

Zhang, Peng, Wang, Xiang, Wang, Futian, Zeng, An, and Xiao, Jinghua (2016b). "Measuring the Robustness of Link Prediction Algorithms Under Noisy Environment". In: *Scientific Reports*. DOI: 10.1038/srep18881.

Zhang, Si, Tong, Hanghang, Xu, Jiejun, and Maciejewski, Ross (2019). "Graph convolutional networks: a comprehensive review". In: *Computational Social Networks* 6.1, pp. 1–23. DOI: https://doi.org/10.1186/s40649-019-0069-y.

Zhang, Yongheng, Lu, Yuliang, Yang, Guozheng, Hou, Dongdong, and Luo, Zhihao (2022). "An Internet-Oriented Multilayer Network Model Characterization and Robustness Analysis Method". In: *Entropy*. DOI: 10.3390/e24081147.

Zhou, Jie, Cui, Ganqu, Hu, Shengding, Zhang, Zhengyan, Yang, Cheng, Liu, Zhiyuan, Wang, Lifeng, Li, Changcheng, and Sun, Maosong (2020). "Graph neural networks: A review of methods and applications". In: *AI open* 1, pp. 57–81. DOI: https://doi.org/10.1016/j.aiopen.2021.01.001.

Zhou, Tao, Lü, Linyuan, and Zhang, Yi-Cheng (2009). "Predicting missing links via local information". In: *The European Physical Journal B* 71, pp. 623–630. DOI: https://doi.org/10.1140/epjb/e2009-00335-8.

Zhu, Tongtian (2020). "Analysis on the applicability of the random forest". In: *Journal of Physics: Conference Series*. Vol. 1607. 1. IOP Publishing, p. 012123. DOI: https://doi.org/10.1088/1742-6596/1607/1/012123.

Zou, Lei and Xi, Xi (2024). "The Influence of Firm Knowledge Characteristics on Technological Innovation: A Multilevel Network Structure Perspective". In: *The Euraseans Journal on Global Socio-Economic Dynamics*. DOI: 10.35678/2539-5645.1(44).2024.130-146.