UNIVERSITÀ DEGLI STUDI DI TRIESTE

UNIVERSITÀ DEGLI STUDI DI TRIESTE

XXXVIII CICLO DEL DOTTORATO DI RICERCA

in

Applied Data Science and Artificial Intelligence

Finanziato dall'Unione europea - NextGenerationEU
Funded by the European Union - NextGenerationEU

Methods for the analysis of complex group formation mechanisms in attributed networks

Leader-influence, homophily, and core–periphery patterns

Settore scientifico-disciplinare: SECS-S/05

Dottoranda:
Sara GEREMIA

Coordinatore:
Prof. Francesco PAULI

Supervisore di tesi:
Prof. Domenico DE STEFANO

Co-supervisore di tesi:
Dr. Michael FOP

ANNO ACCADEMICO 2024/2025

Finanziato dall'Unione europea NextGenerationEU | Ministero dell'Università e della Ricerca | Italiadomani | UNIVERSITÀ DEGLI STUDI DI TRIESTE

Abstract

Community detection has become a fundamental tool for the analysis of complex networks, yet many real-world systems challenge the assumptions underlying traditional approaches. Strong degree heterogeneity, the presence of influential actors, and the availability of node attributes often limit or question the effectiveness of purely topology-based methods, which tend to overlook the mechanisms driving group formation. Addressing these limitations requires frameworks that explicitly account for distinct structural roles and aggregation driven by attribute similarity.

Community detection in heterogeneous and attributed networks can be approached from different perspectives. Methods focused on standard definitions of communities emphasise connectivity patterns but struggle in the presence of hubs and high inter-community connectivity, while attribute-based approaches capture similarity but may underweight structural organisation. This thesis adopts an integrated perspective and treats degree heterogeneity and leadership as informative features rather than distortions. By doing so, the proposed frameworks advance an interpretable view of network organisation that extends beyond individual nodes to explicitly capture community-level structure and hierarchy, highlighting how leadership and connectivity patterns shape the formation, interaction, and positioning of communities within the broader network.

This thesis draws on a diverse set of large-scale relational datasets spanning labour flow, housing, and scientific collaboration networks. Empirical analyses are conducted using regional administrative data on worker mobility, spatially explicit housing market data, European research collaboration data derived from the Horizon 2020 and Horizon Europe programmes, and bibliometric co-authorship data on Italian academic scholars constructed from the Italian Ministry of University and Research (MUR) registry and Scopus publication records.

Three main contributions are presented. First, empirical analyses demonstrate how structurally prominent actors shape group formation across different applied contexts. Second, a density-based community detection framework is extended to integrate structural information and node attributes, allowing homophily to stabilise community boundaries and redefine leadership in terms of attribute-based community representativeness. Finally, a community-level approach to core-periphery detection is introduced, highlighting the role of peripheral community leaders in maintaining connectivity within the overall system.

Acknowledgements

The doctoral scholarship is co-financed with European Union resources, NextGeneration EU- National Recovery and Resilience Plan, Mission 4- Component 1- Investment 4.1–CUP J92B22000900007.

Sara Geremia and Domenico De Stefano acknowledge financial support under the National Recovery and Resilience Plan (NRRP), with European Union resources, NextGeneration EUNational Recovery and Resilience Plan, Mission 4–Component 1–Investment 4.1–Project Title Models and methods for the analysis of collaboration networks– CUP J53D23011540006.

The authors thank Agenzia Lavoro & SviluppoImpresa–Friuli Venezia Giulia for the support.

Contents

Abstract **i**

Acknowledgements **iii**

List of Abbreviations **xvii**

List of Symbols **xxi**

Introduction **1**
Motivations 4
Research contributions 6
Publications 8

I Group formation driven by leader-influence **11**

Introduction **13**

1 Proximity and role-based embeddings in a regional labour flow network **17**
1.1 Motivating example 18
1.2 Methodology 19
1.2.1 Proximity vs. role-based node embeddings 19
1.3 Application: the FVG labour flow network 21
1.4 Final remarks 23

2 Spectral clustering of Berlin's housing spatial network **25**
2.1 Introduction 25
2.2 Literature review 27
2.2.1 Hedonic pricing models 27
2.2.2 Capturing locational effects 27
2.3 Motivating example 30
2.3.1 Descriptive statistics 32
2.4 Methodology 35
2.4.1 Clustering 35
2.4.2 Hedonic pricing models 38
2.5 Application: the Berlin's housing spatial network 39

2.5.1 Clustering . . . 39
2.5.2 Hedonic pricing models . . . 40
2.6 Final Remarks . . . 43

II Group formation driven by leader-influence and homophily 47

Introduction 49
Relevant literature . . . 50

3 Distance-based approach for density-based community detection 53
3.1 Motivating example . . . 53
3.1.1 Structure-driven leader identification . . . 53
3.1.2 The Les Misérables network . . . 55
3.2 Methodology . . . 57
3.2.1 Attribute-driven density-based community detection . . . 57
3.3 Simulation study . . . 59
3.3.1 Simulation design . . . 59
3.3.2 Simulation results . . . 62
3.4 Application: the Les Misérables network . . . 66
3.5 Final remarks . . . 68

4 Attribute-driven density estimation for density-based community detection 71
4.1 Motivating example . . . 71
4.1.1 The hydrogen Horizon network . . . 73
4.2 Methodology . . . 76
4.2.1 Attribute-driven leader identification . . . 76
4.2.2 Attribute-driven density-based community detection . . . 78
4.3 Simulation study . . . 80
4.3.1 Simulation design . . . 80
4.3.2 Simulation results . . . 83
4.4 Application . . . 84
4.4.1 The Les Misérables network . . . 84
4.4.2 The hydrogen Horizon network . . . 86
4.5 Final remarks . . . 93

III Community-level core-periphery patterns 97

Introduction 99
Relevant literature . . . 100

5 Detecting community-level core-periphery structures 103
5.1 Motivating example . . . 103
5.2 Methodology . . . 105
5.2.1 Community-level core-periphery structures . . . 105
5.2.2 Community-level core-periphery classification . . . 107
5.3 Simulation study . . . 110
5.3.1 Simulation design . . . 110
5.3.2 Simulation results: objective function behavior . . . 110
5.3.3 Simulation results: core-periphery classification performance . 111
5.4 Application: Italian academics co-authorship network . . . 114
5.4.1 Communities detected using SBM . . . 114
5.4.2 Communities defined by geographical regions . . . 120
5.5 Final remarks . . . 121

Conclusion 125

A Open Science 129
A.1 Data availability . . . 129
A.2 Code availability . . . 130

B Appendix to Chapter 2 131
B.1 Results for alternative clustering configurations . . . 131
B.2 Additional application: results for Hamburg . . . 131

C Appendix to Chapter 3 133
C.1 Additional results of simulation study . . . 133

D Appendix to Chapter 4 135
D.1 Details of data generation in simulation study . . . 135
D.2 Additional results of simulation study . . . 136

E Appendix to Chapter 5 143
E.1 Details of data generation in simulation study . . . 143
E.2 Alternative summary statistics . . . 144
E.3 Additional results of simulation study . . . 144
E.4 tf-idf distributions . . . 145

Bibliography 149

List of Figures

1.1 Skip-Gram Node2vec and Role2vec architecture. 20
1.2 Visualisation of coreness and strength. Node size represents the strength, while node colour represents universities and public research organisations. 22
1.3 Two-dimensional UMAP visualisation of node embeddings generated from the LFN using (a) Node2vec and (b) Role2Vec. 23
1.4 (a) Elbow method for choosing the optimal number of clusters K. (b) Two-dimensional UMAP visualisation of Role2vec embeddings partitioned into four groups. 24

2.1 Flow chart summarising data cleaning and selection steps used to construct the final sample. 32
2.2 Berlin districts map and 1 km^2 grid level map, with cells coloured according to district membership. 33
2.3 Flats for sale density in $1\,km^2$ cells in Berlin in 2021-2022. 39
2.4 Spatial clustering results using (a) constrained K-means on cell coordinates with $K = 303$, (b) constrained spectral clustering on weighted spatial network with $K = 303$. 40

3.1 Leader-based toy networks with (a) lower, and (b) higher levels of mixing. Node colours indicate density-based clusters, while node size reflects its degree centrality, and node shape its role as community leader or member. 54
3.2 Les Misérables network. (a) Network partition into 20 reference communities derived from characters' first appearance in the novel. Node colours indicate true community membership; node size reflects degree centrality; and node shape represents leaders, defined as the nodes with the maximum degree in their respective communities. (b) Leaders' connectivity matrix: black cells denote the presence of an edge between two leaders in the original network. 56

3.3 Normalised Mutual Information (NMI) for different values of weight α and mixing parameter μ. Performance on the toy network is compared between adjacency-based and Jaccard-based approaches and tested across different attribute types, including Gaussian, Student's t-distributed, binary, and mixed data. High median NMI values (computed over network replicates) are in blue, low values in red. . . . 63
3.4 **Adjacency matrix** – Normalised Mutual Information (NMI) distribution across different numbers of nodes n, numbers of communities K, mixing levels μ, and attribute–structure weights α. Results are shown for simulated networks with uniform community sizes. 64
3.5 **Jaccard matrix** – Normalised Mutual Information (NMI) distribution across different numbers of nodes n, numbers of communities K, mixing levels μ, and attribute–structure weights α. Results are shown for simulated networks with uniform community sizes. 65
3.6 Les Misérables character network. Attributive edges are shown in red and structural edges in grey, prior to their weighted combination. Each block, outlined by blue lines, corresponds to a chapter, and block elements indicate characters who early co-appeared in that chapter. . . 67

4.1 (a) Data points generated using a Gaussian model with 5 components, with colour representing density; (b) Block-diagonal edge probability matrix, with assortative-only blocks representing communities; (c) Block-diagonal adjacency matrix derived from (b). 77
4.2 Block-diagonal adjacency matrix representing a leader-based network partition in 5 communities; (a) assortative-only, 0% mixing; (b) 20% mixing; (c) 40% mixing. 81
4.3 Normalised Mutual Information (NMI) distribution across different network sizes and numbers of communities (K) for uniform community sizes. Results are reported for all competing methods, including AttDeCoDe, DeCoDe, binary stochastic block model (SBM), Spectral Clustering (SC), Covariate-Assisted SC (CASC), and SC on Network-Adjusted Covariates (NAC). 84
4.4 Les Miserables network partition in 20 communities with node size representing degree and node colour representing community membership. 85
4.5 Les Misérables network. (a) Two-dimensional MDS projection of node attributes with lighter colours indicating higher GMM density; (b) normalised distributions of GMM density, node degree, and local density. 86
4.6 Distribution of node density in the hydrogen Horizon network (a) Two-dimensional TSNE projection of node attributes with lighter colours indicating higher kNN density; (b) normalised distributions of kNN density, node degree, and local density. 88

4.7 Hydrogen Horizon network partition in communities using AttDeCoDe with kNN density on the project keyword embeddings. 90
4.8 Block-diagonal adjacency matrix of the hydrogen Horizon network, representing intra- and inter-community links identified by AttDeCoDe. 91

5.1 (a) Example of a network partition in 5 communities with core-periphery roles; (b) Adjacency matrix representing a community partition of the nodes; (c) Adjacency matrix representing a community-level core-periphery partition. Blue colour denotes the periphery, while red colour denotes the core. 106
5.2 Scatter plot of objective function (ϕ) vs. balanced accuracy (BA) scores for core-periphery networks with $n = 1000$ nodes, $K = 20$ communities, and core proportions of 25%, 50%, and 75% (panel from left to right). Blue points represent solutions with more peripheral communities than the true structure, while red points indicate solutions with more core communities. The true solution (BA = 1) is highlighted with a square. 111
5.3 **Uniform** community sizes – Balanced accuracy distribution across different network sizes and numbers of communities (K). Results are shown for different community detection methods, including Louvain, Infomap, and SBM, as well as for the Borgatti and Everett (B&E) approach.112
5.4 **Non-uniform** (Dirichlet-distributed) community sizes – Balanced accuracy distribution across different network sizes and numbers of communities (K). Results are shown for different community detection methods, including Louvain, Infomap, and SBM, as well as for the Borgatti and Everett (B&E) approach. 113
5.5 (a) Co-authorship network partition in communities and their core-periphery organisation. (b) Co-authorship connectivity matrix ($\hat{\mathbf{\Theta}}$) representing the community-level core-periphery partition. Blue square denotes intra-periphery connectivity, red square intra-core connectivity. 115
5.6 (a) Distribution of topic term frequency (tf) by cluster and core-periphery organisation. (b) Distribution of title tf by cluster and core-periphery organisation. 117
5.7 Distribution of journal frequency by cluster and core-periphery organisation.118
5.8 (a) Geographic distribution of authors by cluster and core–periphery organisation.(b) Academic sector distribution of authors by cluster and core–periphery organisation. 119
5.9 Top three leaders' degree and betweenness centrality for core and peripheral clusters. 120
5.10 Regional co-authorship connectivity matrix ($\hat{\Theta}$) illustrating the core-periphery partition of the Italian regions. Blue square denotes intra-periphery connectivity, red square intra-core connectivity. 122

C1 **Adjacency matrix** – Normalized Mutual Information (NMI) distribution across different numbers of nodes n, numbers of communities K, mixing levels μ, and attribute–structure weights α. Results are shown for networks with non-uniform (Dirichlet-distributed) community sizes. 133

C2 **Jaccard matrix** – Normalized Mutual Information (NMI) distribution across different numbers of nodes n, numbers of communities K, mixing levels μ, and attribute–structure weights α. Results are shown for networks with non-uniform (Dirichlet-distributed) community sizes. 134

D1 Normalised Mutual Information (NMI) distribution across different network sizes and numbers of communities (K) for **uniform** community sizes. Results are shown for AttDeCoDe (all density estimators) and DeCoDe. 137

D2 Normalised Mutual Information (NMI) distribution across different network sizes and numbers of communities (K) for **non-uniform** community sizes. Results are reported for all competing methods, including AttDeCoDe, DeCoDe, binary stochastic block model (SBM), Spectral Clustering (SC), Covariate-Assisted Spectral Clustering (CASC), and SC on Network-Adjusted Covariates (NAC). 138

D3 Normalised Mutual Information (NMI) distribution across different network sizes and numbers of communities (K) for **non-uniform** community sizes. Results are shown for AttDeCoDe (all density estimators) and DeCoDe. 139

D4 Adjusted Rand Index (ARI) distribution across different network sizes and numbers of communities (K) for **uniform** community sizes. Results are reported for all competing methods, including AttDeCoDe, DeCoDe, binary stochastic block model (SBM), Spectral Clustering (SC), Covariate-Assisted Spectral Clustering (CASC), and SC on Network-Adjusted Covariates (NAC). 140

D5 Adjusted Rand Index (ARI) distribution across different network sizes and numbers of communities (K) for **non-uniform** community sizes. Results are reported for all competing methods, including AttDeCoDe, DeCoDe, binary stochastic block model (SBM), Spectral Clustering (SC), Covariate-Assisted Spectral Clustering (CASC), and SC on Network-Adjusted Covariates (NAC). 141

E1 Scatter plot of the alternative objective function $\tilde{\phi}(\mathbf{z}; \hat{\mathbf{\Theta}})$ using median and IQR vs. balanced accuracy for core-periphery networks with $n = 1000$ nodes, $K = 20$ communities, and core proportions of 25%, 50%, and 75%. Blue points represent solutions with more peripheral communities than the true structure, while red points indicate solutions with more core communities. The true solution (Balanced Accuracy = 1) is highlighted with a square. 145

E2 (a) F1 score distribution across different network sizes (n) and numbers of communities (K) for uniform community sizes. (b) F1 score distribution across different network sizes (n) and numbers of communities (K) for non-uniform (Dirichlet-distributed) community sizes. 146
E3 (a) Estimated number of clusters distribution across different network sizes (n) and numbers of communities (K) for uniform community sizes. (b) Estimated number of clusters distribution across different network sizes (n) and numbers of communities (K) for non-uniform (Dirichlet-distributed) community sizes. 147
E4 (a) Distribution of topic tf-idf by cluster and core-periphery organisation. (b) Distribution of title tf-idf by cluster and core-periphery organisation. 148

List of Tables

2.1 Overview of variables, definitions, and data sources used in the analysis. 34
2.2 Summary of characteristics for flats listed for sale in Berlin, 2021–2022 (N = 90,849). 35
2.3 Summary of characteristics for 1 km^2 Berlin raster cells in 2017 (N = 611). 36
2.4 Predictive performance of hedonic models using different locational attributes, along with percentage improvements relative to the baseline model without locational controls. 41
2.5 Hedonic price model estimates. 42
2.6 Standardised variable-importance values for all predictors in the hedonic model with spectral clustering. 43

3.1 Performance of ANDeCoDe on Les Misérables benchmark network under different structural similarity measures (adjacency vs. Jaccard), weighting schemes (unweighted vs. weighted networks), and attribute–structure weights α. Bold values indicate the best performance within each block. 68

4.1 Distribution of organisations involved in hydrogen Horizon projects by country, activity type, and dominant research topics. 74
4.2 Les Misérables network. Agreement in terms of Normalised Mutual Information (NMI) and Adjusted Rand Index (ARI) between estimated and early appearance partition, along with estimated number of clusters ($\hat{K}$), across community detection methods. 87
4.3 Composition of the five largest clusters (allowing for equal frequencies) by country and activity type. 91
4.4 Composition of the five largest clusters (allowing for equal frequencies) by dominant research topics. Project keywords associated with cluster leaders in bold. 91

5.1 Co-authorship network characteristics. 104
5.2 Degree distribution by core membership assigned using Borgatti and Everett model. 116
5.3 Frequency of leaders by sector, role, and location across core and peripheral clusters. 121
5.4 Frequency of regional leaders by sector and role across core and peripheral clusters. 122

B.1 Predictive performance of models using locational attributes derived from constrained K-means and spectral clustering across varying mean cluster sizes. 131
B.2 Predictive performance of hedonic models for Hamburg using different locational attributes (K = 70), with percentage changes relative to the baseline model without locational controls. 132

E1 Scaling of edge probabilities by edge type. 144

List of Abbreviations

Chapter 1

LFN	Labour Flow Network
FVG	Friuli Venezia Giulia
PRO	Public Research Organisation
RL	Representation Learning
ISCO	International Standard Classification of Occupations
UMAP	Uniform Manifold Approximation and Projection
UniTS	University of Trieste
UniUD	University of Udine

Chapter 2

HPM	Hedonic Pricing Model
GIS	Geographic Information Systems
SKATER	Spatial 'K'luster Analysis by Tree Edge Removal
OLS	Ordinary Least Squares
RMSE	Root Mean Squared Error
MAE	Mean Absolute Error

Chapter 3

SBM	Stochastic Block Model
DC-SBM	Degree-Corrected Stochastic Block Model
SC	Spectral Clustering
CASC	Covariate-Assisted Spectral Clustering
NAC	Network-Adjusted Covariates
DeCoDe	Density-based Community Detection
ANDeCoDe	Attributed Networks Density-based Community Detection
tf–idf	Term Frequency–Inverse Document Frequency
NMI	Normalized Mutual Information
ARI	Adjusted Rand Index
MAD	Median Absolute Deviation

Chapter 4

EU	European Union
R&D	Research and Development
CORDIS	Community Research and Development Information Service
RoE	Rest of Europe
PRC	Private for-Profit Company
HES	Higher Education Institution
REC	Research Organisation
PUB	Public Organisation
OTH	Other Organisations
EuroSciVoc	European Science Vocabulary
AttDeCoDe	Attribute-driven density for DeCoDe
DeCoDe	Density-based Community Detection
kNN	k-Nearest Neighbours
GMM	Gaussian Mixture Model
DC-SBM	Degree-Corrected Stochastic Block Model
SBM	Stochastic Block Model
SC	Spectral Clustering
CASC	Covariate-Assisted Spectral Clustering
NAC	Network-Adjusted Covariates
ICL	Integrated Classification Likelihood
NMI	Normalised Mutual Information
ARI	Adjusted Rand Index
tf–idf	Term Frequency–Inverse Document Frequency
MDS	Multidimensional Scaling
t-SNE	t-distributed Stochastic Neighbour Embedding
BIC	Bayesian Information Criterion

Chapter 5

SBM	Stochastic Block Model
MUR	Italian Ministry of University and Research
SSD	Scientific Discipline Sector
BA	Balanced Accuracy
F1	F1-score
NMI	Normalised Mutual Information
ARI	Adjusted Rand Index
B&E	Borgatti and Everett core–periphery model
tf	Term Frequency
tf–idf	Term Frequency–Inverse Document Frequency
GA	Genetic Algorithm

List of Symbols

Chapter 1

$G = (V, E)$	Network
V	Set of nodes in the network
n	Number of nodes
i, j	Indices identifying nodes (employers) in the network
v_i	Node i in the network
E	Set of edges in the network
e_{ij}	Edge (or edge weight) between nodes v_i and v_j
p	Index identifying an individual worker
w_{ij}	Weight of the edge between v_i and v_j
w_{ij}^p	Contribution of worker p to the weight of edge e_{ij}
D_i^p	Duration of the latest contract of worker p with employer i
P	number of employees transferred between v_i and v_j
$w_{\max}$	Upper bound on transferable experience between employers
S	Node sequence generated by a random walk
L	Length of a random walk sequence
c_{v_i}	Context (or neighbouring) nodes of node v_i
$\boldsymbol{z}_i$	Embedding vector associated with node v_i
$\boldsymbol{\beta}_j$	Embedding vector of context node v_j
ϕ	Function that maps each node to a role
K	Number of clusters in the K-means algorithm
k	Index identifying clusters
S_k	Within-cluster dispersion for cluster k

Chapter 2

P_i	Sale price of flat i
$X_{s,i}$	Vector of structural attributes for flat i
$X_{n,i}$	Vector of neighbourhood attributes for flat i
$X_{l,i}$	Locational attribute for flat i
β_0	Intercept term in the hedonic price model
β_q	Coefficient associated with covariate X_q
ϵ_i	Error term for flat i
K	Number of spatial clusters
R^2	Coefficient of determination
Δ	Improvement in performance

Chapter 3

$G = (V, E, \boldsymbol{X})$	Node-attributed network
V	Set of nodes in the network
n	Number of nodes
i, j	Indices identifying nodes in the network
v_i	Node i in the network
E	Set of edges in the network
$\boldsymbol{X}$	Node-attribute matrix
$\boldsymbol{x}_i$	Attribute vector associated with node i
e_{ij}	Edge (or edge weight) between nodes i and j
w_{ij}^S	Structural similarity between nodes i and j
w_{ij}^A	Attribute-based similarity between nodes i and j
w_{ij}^α	Combined structural–attribute similarity
α	Weight parameter balancing structure and attributes
$\hat{\delta}(v_i)$	Density estimate of node i
λ	Density threshold
V_λ	Upper-level set of nodes with density $\geq \lambda$
G_λ	Induced subgraph at density level λ
$C(v_i)$	Community membership of node i
$r(v_i)$	Role of node i (core or member)
$\hat{K}$	Number of detected communities
$\tilde{e}_{im}^\alpha$	Maximum-weight edge incident to node i
μ	Mixing parameter controlling inter-community links

Chapter 4

$G = (V, E, \boldsymbol{X})$	Node-attributed network
V	Set of nodes in the network
n	Number of nodes
i, j	Indices identifying nodes in the network
v_i	Node i in the network
E	Set of edges in the network
e_{ij}	Edge between nodes v_i and v_j
$\boldsymbol{X}$	Attribute matrix
$\boldsymbol{x}_i$	Attribute vector of node v_i
$\hat{f}(\cdot)$	Density estimator in attribute space
k	Number of nearest neighbours
$\hat{\gamma}_i$	Gaussian density of node i
$\hat{\delta}(v_i)$	Attribute-driven density of node i
λ	Density threshold
V_λ	Upper-level set of nodes with density $\geq \lambda$
G_λ	Induced subgraph at density level λ
$C(v_i)$	Community assignment of node v_i
$r(v_i)$	Role indicator (core or member)
$\hat{K}$	Number of detected communities
M	Number of mixture components in GMM
p	Dimension of attribute space
μ	Inter-community mixing parameter
d_i	Degree of node i
d_i^{ext}	Number of inter-community links of node i
$\overline{d}_{C(v_i)}$	Mean degree in the community of node i
w_i	Mixing weight for node i

Chapter 5

$G = (V, E)$	Network
V	Set of nodes in the network
E	Set of edges in the network
n	Number of nodes
K	Number of communities
q_k	Community k
Q	Set of communities $\{q_1, \ldots, q_K\}$
n_k	Size of community q_k
m_{kl}	Number of edges between communities q_k and q_l
$\hat{\theta}_{kl}$	Estimated connectivity between communities k and l
$\hat{\mathbf{\Theta}}$	Estimated inter-community connectivity matrix
$\mathbf{z}$	Core–periphery indicator vector for communities
z_k	Core membership indicator for community q_k
$\mathcal{T}_c$	Set of nonzero core–core inter-community connectivities
$\mathcal{T}_p$	Set of nonzero periphery–periphery inter-community connectivities
δ_c	Density of nonzero core–core links
δ_p	Density of nonzero periphery–periphery links
μ_c	Mean core–core connectivity strength
μ_p	Mean periphery–periphery connectivity strength
σ_c	Standard deviation of core–core connectivity
σ_p	Standard deviation of periphery–periphery connectivity
$\phi(\mathbf{z}; \hat{\mathbf{\Theta}})$	Objective function for community-level core–periphery detection
λ	Proportion of communities assigned to the core

Introduction

Community detection is a central problem in network science and has been extensively studied across disciplines. The objective is to identify groups of nodes that are meaningfully related–typically more densely connected or more similar in their connectivity patterns than expected at random–and to use these groups to characterise the organisation of complex networks. Over the past decades, this problem has generated a vast methodological literature, ranging from heuristic and optimisation-based algorithms to statistically principled generative models (see, among others, Newman, 2010; Fortunato and Hric, 2016; Lee and Wilkinson, 2019). A more detailed discussion of these approaches is provided in the individual chapters, where methods are discussed in relation to the specific empirical and methodological settings considered. Throughout this thesis, terms such as clusters, groups, and communities are employed as synonymous, unless otherwise specified.

Despite this abundance of methods, community detection remains an open and evolving research area. Empirical applications repeatedly highlight that community detection is an ambiguous process, and that the inferred community structure is often method-dependent and sensitive to modelling assumptions (Fortunato and Hric, 2016; Lancichinetti and Fortunato, 2011; Morea and De Stefano, 2025). Moreover, many real-world networks exhibit structural features that challenge the core assumptions underlying classical formulations of community structure. These challenges motivate a growing body of work that seeks to move beyond purely topology-based notions of communities, toward frameworks that explicitly model how and why groups form.

A central but often overlooked mechanism shaping community formation in degree-heterogeneous networks is leader influence (Perc, 2014). When connectivity is unevenly distributed, a limited number of nodes accumulate disproportionate prominence and act as focal points for interaction, guiding flows and shaping local organisation. The terms hubs, leaders, influential, focal, anchor or prominent actors are used interchangeably to refer to nodes that occupy structurally central or coordinating positions in the network. In such settings, communities are often better understood as groups organised around influential or representative actors rather than as uniformly dense regions of the network.

This leader-based view contrasts with common formulations of community detection, which implicitly assume relatively homogeneous connectivity within groups. A large body of methodological literature has since relaxed this assumption, recognising that real communities often accommodate internal heterogeneity, with highly connected hubs coexisting alongside more peripheral nodes. In parametric settings, this insight is formalised by the degree-corrected stochastic block model

(Karrer and Newman, 2011), which explicitly allows nodes within the same community to have different expected degrees, and by its extensions, including the popularity-adjusted block model, designed to capture variations in node popularity across communities (Sengupta and Chen, 2018; Noroozi et al., 2021). Hierarchical extensions allow for nested block structures, capturing multi-scale organisation driven by heterogeneous influence (Peixoto, 2014, 2019; Legramanti et al., 2022). These models demonstrate that taking into account degree heterogeneity is essential to avoid mixing up hubs with communities.

Implicit in this observation is an assumption about the type of similarity that is relevant for node clustering–an assumption that can be made explicit through the distinction between role-based equivalence and structural equivalence (White and Reitz, 1983; Lorrain and White, 1971; Rossi et al., 2020). Role-based equivalence identifies nodes as similar when they play comparable roles in the network, even if they connect to different sets of neighbours. From the standpoint of community identification, this shifts the focus from where nodes are connected to how they are positioned. Leaders defined through role-based equivalence are not necessarily the most locally connected nodes, but rather those occupying functionally prominent positions–such as hubs, brokers, or bridges–that shape connectivity patterns across the network. This view is embedded in non-assortative or core-periphery block structures (Borgatti and Everett, 2000; Gallagher et al., 2021; Yanchenko and Sengupta, 2023), role-based embeddings (Rossi and Ahmed, 2014; Ahmed et al., 2018), and positional analyses, in which leadership reflects functional influence rather than neighbourhood overlap.

Structural equivalence, by contrast, identifies nodes as similar when they share the same neighbours or display strong overlap in direct connectivity. When grouping is based on such proximity-driven notions, communities correspond to tightly connected neighbourhoods, and leaders are selected as nodes that dominate local interaction patterns. This perspective underlies assortative block models (Zhang and Peixoto, 2020), proximity-based embeddings (Grover and Leskovec, 2016), and many centrality-driven leader-first algorithms, in which leadership reflects being embedded in a dense local cluster (Menardi and De Stefano, 2022).

Leader-first approaches to community detection can be viewed as an algorithmic response to the same empirical problem addressed by degree-corrected block models. Rather than adjusting for degree heterogeneity, they treat heterogeneity as a generative force in community formation. These methods formalise the intuition that groups emerge around prominent nodes by decomposing the clustering process into two conceptually distinct steps. First, leaders are identified based on a definition of prominence or attraction. Second, other nodes are assigned to leaders, typically by following structural paths, similarity relations, or density-based criteria (Yakoubi and Kanawati, 2014; Helal et al., 2017; Ahajjam et al., 2018; Lu, 2021; Akachar et al., 2025). The definition adopted in the first step is crucial, as it determines which nodes are considered potential anchors and, consequently, the nature of the communities that emerge.

Leader identification requires specifying what it means for a node to be influential or representative. Many approaches define leaders through structural prominence, using measures such as degree, centrality, or local density (Lu, 2021; Menardi and De Stefano, 2022). This perspective treats leadership as an endogenous property of the network topology, arising from heterogeneous connectivity patterns and preferential attachment mechanisms.

Beyond purely structural definitions, attribute-driven homophily introduces an additional, complementary basis for leader identification (McPherson et al., 2001). When nodes preferentially connect to others with similar observed characteristics–such as thematic focus, geographic location, or socio-demographic traits–leadership can be interpreted in terms of representativeness rather than connectivity alone. This perspective extends density-based community detection by redefining attraction in terms of similarity rather than raw connectivity. In doing so, it distinguishes between endogenous popularity, driven by preferential attachment mechanisms such as degree or visibility, and exogenous popularity, which emerges from representativeness in the attribute space. It also aligns with popularity–similarity models of network formation (Papadopoulos et al., 2012, 2014), in which connectivity emerges from the joint effects of structural prominence and attribute affinity.

This shift has motivated a growing literature on community detection in attributed networks, which aims to integrate topological structure with node attributes to improve both interpretability and robustness. Existing approaches include probabilistic models that incorporate covariates into block structures (Stanley et al., 2019), spectral methods that regularise graph partitions using attribute similarity (Binkiewicz et al., 2017; Hu and Wang, 2024), and algorithmic frameworks that combine topology and attributes through weighted similarity measures or joint optimisation criteria (see Chunaev, 2020, for an overview). These methods demonstrate that attributes can complement the network structure, improving the determination of community boundaries.

Overall, the literature indicates that community structure in complex networks emerges from the interplay of multiple mechanisms, including leader-driven attraction, homophily on observed attributes, and higher-order organisation based on role-based equivalence. Although each of these mechanisms has been extensively studied in isolation, relatively few contributions offer unified frameworks that integrate them in an interpretable manner, particularly in empirical settings characterised by sparsity, pronounced degree heterogeneity, and the availability of node attributes. Most work that considers mechanism interplay remains anchored to a density-based view of communities and focuses primarily on disentangling triadic closure from homophily, treating hubs and leadership as secondary or implicit effects (Asikainen et al., 2020; Peixoto, 2022). At the same time, recent evidence suggests that degree heterogeneity—through the emergence of hubs and bridge actors—plays a key role in shaping meso-scale organisation, interacting with homophily in ways that can either reinforce or blur community boundaries (Bachmann et al., 2025; Fuhse and Gondal, 2024; Kozitsin et al., 2023). This gap is particularly relevant for sparse collaboration

and interaction networks, where a small number of highly connected actors can dominate connectivity patterns and where attributes provide essential context for interpreting leadership and group membership.

Grounded in applications to regional innovation and economic networks, spatial systems, and scientific collaboration, the approaches developed in this thesis show how explicitly accounting for these group-formation mechanisms yields insights that are not attainable with methods relying solely on network topology or on attributes considered in isolation.

Motivations

Community structure is a pervasive feature of relational data, but in many applied contexts, the questions that motivate network analysis go beyond simple "cluster search". Rather, practitioners are often interested in understanding why clusters form, which actors anchor them, and how clustering results translate into interpretable and policy-relevant units. This thesis is motivated by empirical contexts where networks encode meaningful interactions–worker mobility, spatial housing markets, and scientific collaboration–yet exhibit structural complexities that make standard community detection either uninformative or sensitive to small changes in the data or in the modelling choices, thus unreliable. Across these applications, two recurring challenges emerge: (i) pronounced heterogeneity in connectivity, which concentrates influence in a limited set of leaders and can distort purely topology-driven partitions; and (ii) exogenous constraints and node attributes that guide attachment and may be integrated with topology to recover homophilic groups.

The first empirical setting concerns *labour flow networks* constructed from employment contract histories in Friuli Venezia Giulia (FVG). Using administrative records organised as contract "events" between 2014 and 2021, we reconstruct worker transitions between employers (regional-based organisations) and represent them as a directed, weighted network in which edges measure mobility and knowledge transfer. The resulting graph is sparse and, by construction, restricted to high-skill occupations in science, engineering, and information and communication technology, a choice motivated by the aim of isolating innovation-relevant knowledge flows among firms, universities, and public research organisations. This setting motivates community detection for two reasons. First, clustering employers based on mobility patterns offers an empirical lens on the organisation of regional labour markets, potentially revealing functional labour-market segments and the role of innovation actors. Second, the network is intrinsically degree-heterogeneous: a small number of employers mediate disproportionate inflows and outflows, reflecting economic centrality. At the same time, mobility is strongly shaped by geography, so proximity-driven ties risk producing trivial geographic partitions. These features motivate methods that separate where an employer is from what role it plays, and that treat

degree heterogeneity as information about prominence and leadership within the labour system.

The second empirical setting is related to *spatial organisation in housing markets*, where the goal is to capture locational value when fine-grained spatial data are limited. We analyse flat listings for sale in Berlin in 2021–2022 from ImmoScout24 (Schaff and Thiel, 2023), enriched with neighbourhood covariates. Berlin provides a particularly demanding environment: historical institutional change, privatisation, uneven redevelopment, and segmented neighbourhood trajectories create strong spatial heterogeneity in price formation. Hedonic pricing models are well-suited to decomposing prices into attribute contributions, but they struggle to represent location effects when spatial identifiers are coarse and neighbourhood boundaries are not known a priori. Here, community detection plays a different role: it provides a data-driven definition of neighbourhoods that can be integrated as location effects in pricing models. Crucially, the spatial units (grid cells) differ markedly in housing density, so any meaningful partition must adapt to heterogeneous local prominence–dense micro-areas should be allowed to form small clusters (or remain isolated), whereas sparse areas should aggregate into broader neighbourhoods.

While these first two studies demonstrate that leaders and density heterogeneity can guide node clustering in practice, structural prominence does not always coincide with community representativeness. This becomes explicit in the motivating examples in Part II. In density-based and leader-first approaches, leaders are typically identified as peaks in a node-wise density landscape, often proxied by degree or other centrality measures. The toy example illustrates how leaders can become structural connectors that erode cluster separability, while the Les Misérables benchmark shows the same phenomenon in an empirical network with strong degree heterogeneity: central characters can dominate clustering based solely on topology, producing no meaningful partitions even when a reference grouping exists. At the same time, many real-world networks provide rich node information that capture homophily and can stabilise community boundaries when topology alone becomes ambiguous. These observations motivate the development of methods that integrate the influence of leaders with attribute-based similarity, restoring interpretability by aligning attraction with representativeness rather than overall popularity.

A closely related motivation comes from *EU-funded collaboration networks* derived from CORDIS data on Horizon projects (European Commission, 2024), where the community structure reflects both relational ties and thematic organisation. In this context, nodes (organizations) participate in projects that report thematic descriptors, so the same bipartite events generate both the collaboration graph and a textual attribute space. Organisations involved in many projects become structurally popular and simultaneously accumulate broader thematic profiles, creating a coupling between popularity and similarity that may play a central role in network growth. Hydrogen projects are a particularly illustrative case because they are explicitly designed to connect heterogeneous actors–research institutes, companies, and public institutions–in regional innovation ecosystems. This motivates a perspective of

community identification in which "leaders" are not only structurally central but also thematically representative, and in which communities emerge around dense regions of the attribute space while remaining constrained by observed ties of collaboration.

Finally, the last part of the thesis is motivated by a shift from node-level centrality to group-level structural roles in scientific collaboration. Using a *co-authorship network of Italian academics* (2012–2022), enriched with information on discipline, location, academic role, and research themes, we observe both community structure and high degree heterogeneity. In particular, senior scholars and statisticians often bridge across otherwise separate research groups. This suggests that communities may not be equally integrated into the broader system: some may function as central subsystems that connect widely, while others remain relatively isolated and interact mainly with the core. These observations motivate a shift from detecting communities to characterising their hierarchical organisation–specifically, assessing whether communities themselves exhibit core–periphery patterns. Importantly, this also changes the perspective on leader identification: leaders can be classified by the structural role they play depending on whether they belong to core or peripheral communities, distinguishing actors whose influence is system-wide from those whose prominence is local and may mediate the integration of peripheral groups into the broader collaboration network.

Taken together, the empirical settings provide concrete examples where standard community detection approaches–whether purely topology-based or purely attribute-driven–are insufficient, thereby motivating the methodological developments that follow.

Research contributions

This section presents the thesis structure and highlights its key contributions.

Part I consists of two empirical studies showing how explicitly accounting for heterogeneity in node connectivity and spatial concentration improves both the reliability of the results–namely, their robustness to small changes in the data–and their interpretability. In Chapter 1, we introduce a graph representation learning perspective for regional labour flow networks, comparing proximity-based node embeddings with role-based ones. Proximity-based embeddings are driven by shared neighbours and the intensity of their connections, whereas role-based embeddings abstract from local connectivity patterns to capture nodes' structural roles in the network, such as hubs or bridges. This reveals that universities and research organisations in the data exhibit non-overlapping sets of labour-flow connections (low structural equivalence) while sharing a similar relational role in the overall labour system (high role-based equivalence). From a political perspective, interventions targeting these institutions allow access to distinct segments of the regional economy while exploiting their shared core role.

In Chapter 2, we address both the limited availability of high-resolution spatial information and heterogeneity in spatial distribution by representing space as a weighted network (Anselin, 2024), where link weights are inversely proportional to the density of the connected cells. In this context, degree heterogeneity and weighted connectivity become resources rather than hindering characteristics of complex networks. Treating density as a structural feature and explicitly encoding it in the network weights allows the clustering procedure to recognise heterogeneous levels of spatial centrality, thereby identifying local core cells that inform meaningful spatial segmentation. The resulting connectivity structure is exploited through spectral clustering, which groups sparse areas into larger clusters while allowing dense areas to form smaller clusters, including isolated individual cells. This procedure yields spatially contiguous clusters that adapt to local density variations and can be directly incorporated into pricing models when spatial information is available at an aggregated resolution. The resulting model improves predictive accuracy compared to models based on administrative boundaries and alternative clustering approaches that rely exclusively on georeferenced coordinates.

Part II moves from empirical evidence of leader-influenced structures to methodological contributions in community detection. The analysis is explicitly framed within a density-based community detection (DeCoDe) perspective, in which communities are understood to form around nodes exerting strong local attraction, typically associated with high local density or representativeness. Within this framework, Part II addresses a central question emerging from the empirical results: how the presence of leaders shapes cluster formation, and to what extent leader-driven attraction alone can account for observed community structures.

A key contribution of Part II is to show that leader-based attraction, while central, is often insufficient to explain community formation in real-world networks. The analysis identifies homophily as an additional and complementary mechanism guiding aggregation, whereby nodes preferentially connect to others with similar attributes. While both mechanisms are individually well studied, their joint role has rarely been formalised within a unified community detection framework. Part II addresses this gap by developing and analysing methods that integrate density-based attraction and attribute-driven homophily, demonstrating how their interaction jointly determines community structure in attributed networks.

Chapter 3 contributes a systematic assessment of how to combine topology and attributes in the DeCoDe framework Menardi and De Stefano (2022). A second contribution is to document the importance of the choice of structural similarity measure in the clustering process. Together, the results clarify when purely structural or purely attribute-driven strategies fail, and they motivate principled hybrid designs for attribute-augmented DeCoDe.

Chapter 4 introduces a second attribute-based extension of DeCoDe that shifts the leader-identification step from structural prominence (e.g. degree) to high density in attribute space. The main novelty is conceptual and methodological: communities

are organised around attribute-representative modal actors, allowing leader influence to emerge from covariate concentration rather than from structural dominance. We further propose a simulation framework that extends the degree-corrected stochastic block model by coupling block structure with attribute-driven degree heterogeneity, thereby providing a representative setting to evaluate the proposed methods under realistic sparsity and mixing.

Part III extends the analysis of degree heterogeneity and leader influence from the node level to group-level structural roles. Building on the idea that highly connected or influential nodes shape local organisation, Chapter 5 introduces the first explicit framework for detecting nested core–periphery structures at the community level. Communities are distinguished according to the density and strength of their inter-community connections, capturing how concentrated connectivity and leader-driven interactions aggregate into densely interconnected core communities and more weakly integrated peripheral ones. Importantly, this framework distinguishes leaders by their membership in core or peripheral communities, thereby revealing whether their influence is structurally central to the system or operates from the margins, potentially serving as a bridge for integrating peripheral clusters into the broader network. The methodological contribution consists of an interpretable, flexible, and scalable objective-function-based formulation, which allows the identification of both densely interconnected core communities and peripheral communities that may be internally cohesive yet weakly integrated into the broader system.

Overall, the thesis advances a unified view of network heterogeneity–across roles, homophily, and hierarchical organisation–showing how incorporating these mechanisms yields methods and empirical insights that are not accessible through traditional approaches.

Publications

All the contributions presented in this thesis have either been published in conference proceedings or are currently under review in high-impact journals. I would like to extend my sincere thanks to all the co-authors listed below, whose collaboration and support were essential to the completion of this work.

In particular:

- Chapter 1 is based on [1], co-authored with Fabio Morea (Area Science Park, Trieste) and Domenico De Stefano (University of Trieste).
- Chapter 2 is based on [2], co-authored with Alessio Sardo and Gianluca Cerruti, both from the University of Genova.
- Chapter 3 is based on [3], co-authored with Domenico De Stefano.

- Chapter 4 is based on [4], co-authored with Domenico De Stefano and Michael Fop (University College Dublin).
- Finally, Chapter 5 is based on [5], co-authored with Domenico De Stefano and Michael Fop. The ideas developed in this chapter also informed the work presented in [6] and [7].

[1] Geremia, S., Morea, F., De Stefano, D.: Visualization of Proximity and Role-Based Embeddings in a Regional Labor Flow Network. In: Scientific Meeting of the Classification and Data Analysis Group of the Italian Statistical Society, pp. 113–121. Springer Nature Switzerland (2023).

[2] Geremia, S., Cerruti, G., Sardo, A.: Spectral Clustering of Berlin's Housing Spatial Network to Capture Locational Effects on Pricing. Accepted for publication in *Regional Science, Regional Studies* (2025).

[3] Geremia, S., De Stefano, D.: Density-based community detection combining structure and attribute information. In: Scientific Meeting of the Classification and Data Analysis Group of the Italian Statistical Society, pp. 192–201. Springer Nature Switzerland (2025). ⋆ *Best Paper Award*.

[4] Geremia, S., Fop, M., De Stefano, D.: A density-based framework for community detection in attributed networks. Under review in *Journal of the Royal Statistical Society: Series A*. arXiv preprint (2025).

[5] Geremia, S., De Stefano, D., Fop, M.: Community-level core–periphery structures in collaboration networks. Under review in *Social Networks*. arXiv preprint arXiv:2511.19305 (2025).

[6] Geremia, S., Sardo, A.: Investigating the Lisbon Treaty's Impact on the Co-citation Network of the Court of Justice of the European Union Case Law. In: Scientific Meeting of the Italian Statistical Society, pp. 296–301. Springer Nature Switzerland (2025).

[7] Geremia, S., Sardo, A.: Symbolic vs. Structural Precedent after Lisbon: Co-Citation Networks Analysis of CJEU Case Law (2008–2015). Under review in *Statistical Methods & Applications* (2025).

Part I

Group formation driven by leader-influence

Introduction

Understanding how groups emerge in complex systems requires careful attention to the structural mechanisms that govern how nodes interact, cluster, and diverge. A central mechanism–particularly relevant in networks characterised by heterogeneous distributions of connection frequency and intensity–is leader influence: the tendency of structurally prominent nodes to guide flows and shape local organisation. This part of the thesis investigates how leaders influence group formation in two distinct yet conceptually connected settings: labour flow networks, where a small number of employers dominate worker mobility, and spatial housing networks, where dense urban cells emerge as structurally prominent nodes under spatial contiguity constraints.

Although the two chapters differ in empirical context and analytical tools, they are linked by two core aspects that motivate their inclusion in Part I.

Influence of geographic location in network formation. In both cases, geographical information is central–either as a latent driver of interactions or as an explicit constraint governing network construction.

Chapter 1 examines workers' mobility. Employee mobility across workplaces and municipalities is strongly shaped by geography. The influence of node location on labour flows is well-documented in regional and urban economics, where job transition patterns reflect not only job accessibility, worker specialisation and the attractiveness of hub firms but also workers' residential proximity constraints, commuting costs, and transport connectivity. Because of this, networks of labour flows often exhibit proximity-driven similarity: geographically close firms tend to exchange workers more frequently. This spatial influence has a major analytical implication: direct interactions may reveal only trivial partitions. When geography dominates connections, classical community detection tends to reproduce spatial adjacency rather than uncover informative structural relations. The first chapter applies an embedding-based methodology to move beyond mere proximity clusters by comparing nodes by their structural roles to reveal meaningful labour market structures.

In the second chapter, geographical information enters not as an external influence but as a defining constraint in network and community formation. The network is built from geographically adjacent spatial cells; therefore, contiguity determines edges. This design directly reflects the organisation of urban housing markets, in which local price formation is driven by neighbourhood effects, micro-location attributes,

and path-dependent development patterns. Here, enforcing spatial contiguity in the construction of the network has a methodological advantage: it yields clusters that correspond to actual neighbourhood structures, avoiding the unrealistic spatial fragmentation that often arises from using clustering methods that do not account for geographical location. In this sense, geography is again a structural feature that guides the detection of meaningful groups.

Influence of degree heterogeneity in group formation. A second shared characteristic provides the conceptual heart of Part I: both networks exhibit degree heterogeneity, a feature that plays a central role in shaping group formation.

Workplaces differ substantially in the amount of labour inflow and outflow they mediate. Such heterogeneity reflects economic centrality, industrial composition, accessibility, and organisational roles within the broader labour system. The methodological strategy adopted in the first chapter addresses degree heterogeneity explicitly. By using role-based embeddings, the analysis characterises nodes in terms of their structural functions: attractors of mobility, transit nodes, local hubs, and peripheral units. This approach has two advantages:

1. It captures structurally important nodes that act as leaders within the labour flow system.

2. It reveals patterns obscured by raw interaction data, disentangling structural roles from mere geographic adjacency.

With regards to Chapter 2, the examined Berlin housing network is constructed from spatial contiguity, and we also have information on the density of flats per cell–a proxy for the structural prominence of a location in the urban fabric. Because property-level data are geographically aggregated, this cell-level density acts as a measure of node degree heterogeneity: densely populated cells represent representative micro-areas which can be considered different from surrounding areas, while low-density cells are allowed to cluster with close cells.

In the analysis, this heterogeneity becomes a resource rather than a limitation. Treating density as a structural feature allows the clustering procedure to recognise diverse centrality levels across the network, thereby identifying local core cells that inform spatial segmentation.

Taken together, the two chapters demonstrate how leader-influence mechanisms operate in practice across distinct settings:

- In labour mobility networks, leaders emerge as structurally prominent nodes whose flow patterns define labour market regions beyond geography.

- In spatial housing networks, densely populated or structurally central cells shape the boundaries and internal composition of neighbourhood clusters.

In both cases, identifying structurally important nodes is essential to uncovering the formation of groups.

This empirical foundation sets the stage for the developments in Part II, where leader influence is combined with homophily to explain aggregation in attributed networks, and Part III, which extends the focus from node-level roles to community-level core–periphery structures, which themselves depend on the presence of leader actors.

Chapter 1

Proximity and role-based embeddings in a regional labour flow network

The mobility of workers creates a network of connections that reflects the interconnectivity between employers, the so-called Labour Flow Network (LFN). Such network data can reveal insights into the structure of the relationships between employers, which can be used to identify communities of employers that share geographic location, industry, and workforce characteristics (Park et al., 2019). The mobility of individual workers, from one job to another, plays a role in knowledge spillover within economies (Eriksson, 2011). The analysis of LFNs' topological features and their community structures shows that LFNs can be useful in identifying influential firms (i.e., employers), including those with relatively high turnover that nonetheless act as hubs for skilled labour and innovation (Guerrero and Axtell, 2013). In such cases, high turnover may signal active talent renewal and strong demand for the firm's positions. Therefore, examining similar connectivity patterns in such networks is a key step in understanding the role and position of some targeted employers in the labour market. The role of large public sector organisations, such as universities, in the economic context under study can be determined by their centrality and relationship with industries employing a high number of experienced professionals (Smallbone et al., 2015).

This work aims to investigate the structure of an LFN in Friuli Venezia Giulia (FVG), a medium-sized region in northern Italy ranked as one of the leading regions for innovation. FVG stands out as one of the few regions in Italy classified as a strong innovator, with a demonstrated increase in innovation performance over time (European Commission et al., 2023). We construct the LFN based on a dataset provided by the Regional Observatory on Policies and the Labour Market of FVG, which encodes the commencement or termination of employment contracts as single "events", as outlined in Morea and De Stefano (2023). In particular, the data are filtered to include only a subset of professional groups, encompassing highly skilled workers whose expertise is potentially related to the most innovative sectors in the region.

Previous studies on FVG have shown that knowledge and innovation diffusion are guided by the role of academia and the leading Public Research Organizations

(PROs) operating at the local level (De Stefano and Zaccarin, 2013). For this reason, we are especially interested in exploring the structure of such a peculiar regional LFN, as well as the position and role of regional universities and prominent PROs.

Understanding the structure of such LFN involves dealing with graph data containing rich relational information. Traditional machine learning algorithms require hand-engineered feature representation, which is labour-intensive and relies on domain-specific knowledge. Representation Learning (RL) provides an alternative learning approach to automatically represent graph data using low-dimensional vectors (Cui et al., 2022). The learned embeddings can be used with data visualisation and clustering techniques to generate representations of graphs useful for discovering communities, hub nodes, and other hidden structures.

The graph RL task can be performed to assess the potential of universities and research institutions as drivers of economic development and innovation in FVG. The interest is in investigating whether exploring the network by examining both relational proximity and recurring structural patterns yields valuable insights. In this paper, we focus on the role and position of specific regional organisations, namely universities and leading PROs, according to two different node equivalence notions. The two graph RL we adopted are well-suited for such tasks. The identification of groups of equivalent nodes could also be performed using other approaches, such as generalised blockmodeling (Doreian et al., 2007). However, this approach allows for a global positional analysis and can be problematic for large and dense networks.

1.1 Motivating example

The present study utilises employment contract data in FVG for constructing an LFN, wherein nodes represent employers and edges between two nodes v_i and v_j signify the transition of at least an employee p from employer v_i to employer v_j. The employment contract data are organised as 'events', where each event represents the initiation or termination of an employment contract. The dataset is anonymised and includes 1,155,342 employment events from 74,317 local units of companies, universities, and PROs in the FVG region, between 2014 and 2021. For the purpose of this paper, the data are filtered out to include only a subset of professional groups, namely ISCO-21 (science and engineering occupations) and ISCO-25 (information and communication technology occupations), as defined by the International Standard Classification of Occupations 2008 (European Union, 2009). The dataset comprises approximately 60,164 events involving 1,890 employers and 16,474 employees. To ensure data reliability, the raw dataset is cleaned and aggregated (e.g., adding implicit contract terminations, identifying actual workplaces for employment agencies, and grouping professional profiles).

We then constructed a network of 'transitions,' defined as a cessation event followed by a start event involving the same employee and two different employers. Transitions involving the same employer or those not fully traceable within the region,

such as self-employment, leaving the labour force, or finding a new job in a different region, were excluded. The final network contains 1,084 nodes (unique employers involved in regional transitions) and 1,641 edges (identifiable transitions), weighted by the number of transitions and the relevance of each transition. Specifically, the edge weight accounts for the knowledge spillover effect: it is proportional to the duration of the experience gained by employee p while working for employer v_i and transferred to employer v_j. The weight of this transfer is given by $w_{ij}^p = min(D_i^p, D_j^p, w_{max})$, where D_i^p is the duration of the contracts of p with v_i and D_j^p is the duration of the contracts of p with v_j (both expressed in years or fraction of years). The parameter w_{max} sets the maximum contribution that an individual's experience on a given link can make to the edge weight, capping the weight at the equivalent of 5 years of employment. The edge weight w_{ij} is given by $w_{ij} = \sum_{p=1}^{P} w_{ij}^p$, where P is the number of employees transferred between v_i and v_j.

1.2 Methodology

1.2.1 Proximity vs. role-based node embeddings

In this work, two methods for graph RL are employed: Node2Vec (Grover and Leskovec, 2016) and Role2Vec (Ahmed et al., 2018). The methods differ in their approaches to preserving the graph's structure, as they are based on two distinct definitions of node structural similarity, known as structural equivalence and role-based equivalence (White and Reitz, 1983; Lorrain and White, 1971; Rossi et al., 2020), respectively. Two nodes are structurally equivalent if they are relationally close: structural equivalence captures similarities in the way nodes are connected to the same neighbouring nodes. Two nodes are structurally similar if they occupy similar roles or positions in the network: role-based equivalence captures similarities in the way nodes contribute to the overall structure of the network. Understanding these differences is crucial because the choice between structural and role-based equivalence reflects the emphasis placed on different aspects of the graph topology.

Node2vec might be more suited when the goal is to capture similarities in how nodes interact with each other. It is designed to map structurally similar nodes to similar proximity-based embeddings, considering both local and global structural information. This is achieved through the use of a random walk strategy, generating sequences of L nodes ($S : v_1, v_2, ..., v_L$) that capture neighbourhood relationships. Node2Vec shares the Skip-Gram model architecture (Figure 1.1) with the widely known Word2Vec (Mikolov et al., 2013) and optimises embeddings so that, given a node, it can predict its neighbours. The objective is to maximize the log-probability of observing a network neighbourhood for each node (Grover and Leskovec, 2016): $max \sum_{v_i \in V} \sum_{v_j \in c_{v_i}} \log Pr[v_j|v_i]$, where c_{v_i} represents v_i's context nodes, i.e., nodes that are topologically related to v_i. The conditional probability $P[v_j|v_i]$ is often modelled using a softmax function: $P[v_j|v_i] = \frac{\exp(\boldsymbol{z_i} \cdot \boldsymbol{\beta_j})}{\sum_{l=1}^{n} \exp(\boldsymbol{z_i} \cdot \boldsymbol{z_l})}$, where $\boldsymbol{z_i}$ denotes v_i embedding and

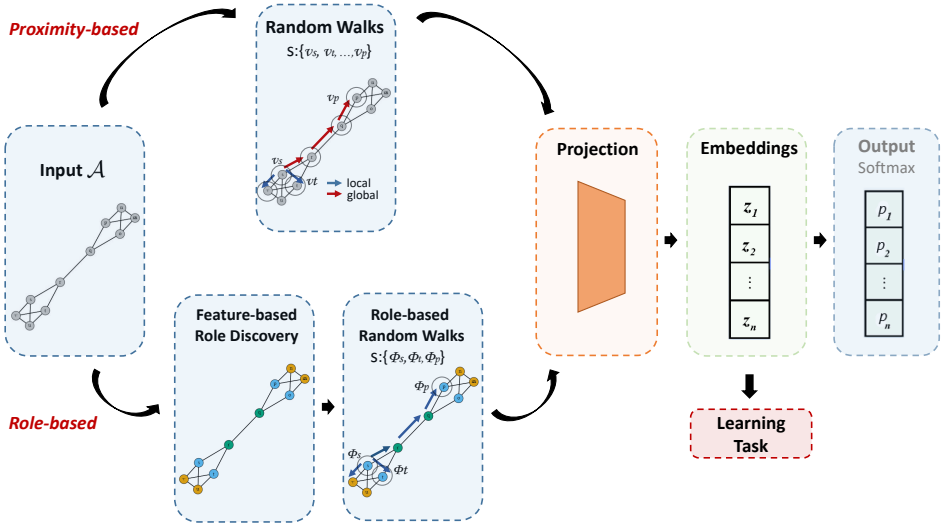


FIGURE 1.1: Skip-Gram Node2vec and Role2vec architecture.

$\boldsymbol{\beta_j}$ the embedding of context node v_j. The node2vec embedding dimension is set to 50, providing a good trade-off between capturing structural information and limiting model complexity, in line with common practice in the literature.

Role2Vec is designed so that nodes sharing similar structural roles–defined by their connectivity patterns–are mapped to similar latent representations, even if they lie in different regions of the graph (Rossi and Ahmed, 2014; Rozemberczki, 2019). The latent representations are learned using a feature-based random walk approach, where walks find similar nodes identified by structural properties and higher-order graph features (e.g., hubs, triangles, 4-cycles, etc.). A feature-based random walk is a random sequence of adjacent nodes sharing the same role, $S : \phi_{v_1}, \phi_{v_2}, ..., \phi_{v_L}$, with ϕ a function that maps each node to a role. S is induced by a sequence of indices $(v_1, v_2, ..., v_L)$ generated by a random walk. Role2vec uses the generated sequences to train a Skip-Gram model and learn 50-dimensional role-based embeddings. The approach can be viewed as incorporating a relaxed, embedding-based form of global regular-equivalence information: rather than enforcing strict recursive equivalence conditions on the graph, it learns continuous role representations that approximate nodes sharing similar patterns of ties to other roles, thereby providing a data-driven relaxation of regular equivalence.

Both Node2vec and Role2Vec have been shown to achieve state-of-the-art performance on various graph RL tasks, but the choice of method depends on the nature of the dataset and the task at hand. The performance evaluation in this work is conducted without true labels for a supervision task; thus, it is based on visualisation and cluster analysis. We applied the popular K-means clustering approach to the 50-dimensional embeddings and used the within-cluster dispersion (S_k) to assign clustering scores and select the optimal number of clusters K. The results of the

models are evaluated by exploring the network latent representations re-embedded with Uniform Manifold Approximation and Projection (UMAP) (McInnes et al., 2018). UMAP is a dimensionality reduction technique that takes local structure into account to increase the data representation quality in terms of clusterability.

For the analyses, we used the `role2vec` (Rozemberczki, 2019) and `node2vec` (Cohen, 2019) Python packages to generate the embeddings, the UMAP package to reduce the embedding dimensions for visualisation purposes, and the `ggplot2` (Wickham, 2016) R package for creating the visualisations.

1.3 Application: the FVG labour flow network

Centrality measures are computed for each node in the network to compare employers and reveal their relative positions within the broader knowledge-transfer system. This allows one to track how different organisations contribute to and benefit from labour-driven knowledge spillovers.

The strength of a node is the sum of the weights of the edges incident to that node: $s_i = \sum_{j \in N(v_i)} w_{ij}$, where $N(v_i)$ is the set of neighbours of v_i and w_{ij} represents the weight of the edge between nodes v_i and v_j. In this context, the strength reflects the total amount of experience-based knowledge spillover that an employer receives from or transfers to other organisations over the observation period.

The coreness of a node is a measure of the node's position within the network's hierarchical structure, based on its connectivity. For example, a node has a coreness of 5 if it belongs to the 5-core of the network. The 5-core is a maximal subgraph in which every node is connected to at least 5 other nodes. In this context, coreness can be interpreted as the capacity of an organisation to participate in dense, mutually reinforcing knowledge-transfer clusters.

In Figure 1.2, the y-axis displays the logarithmic values of the node strength. Research and higher education institutes, i.e., Universities of Trieste (UniTS) and Udine (UniUD), SISSA, OGS and Elettra, are highlighted with colours as they represent the largest research institutions in terms of both employee count and their impact on the regional economic system SiS FVG (2024). The scatterplot shows that such institutions are among the nodes with the highest coreness and strength. The grey nodes with high strength and coreness correspond to the main industrial companies in FVG (Danieli, Fincantieri, Electrolux, etc.), which have a high volume of employment.

Figure 1.3 compares the UMAP visualisations of node embeddings learned with the methods. The distances between nodes in the embedding space reflect structural (left panel) and role-based (right panel) equivalence in the original graph. The size of the nodes indicates the strength, thus, employers employing a high number of experienced professionals are shown with bigger nodes. The colour of the nodes highlights universities and PROs in the network. The left plot seems to suggest the absence of a community structure: coloured and large-sized nodes are scattered in distinct regions of the graph, with no discernible patterns. From

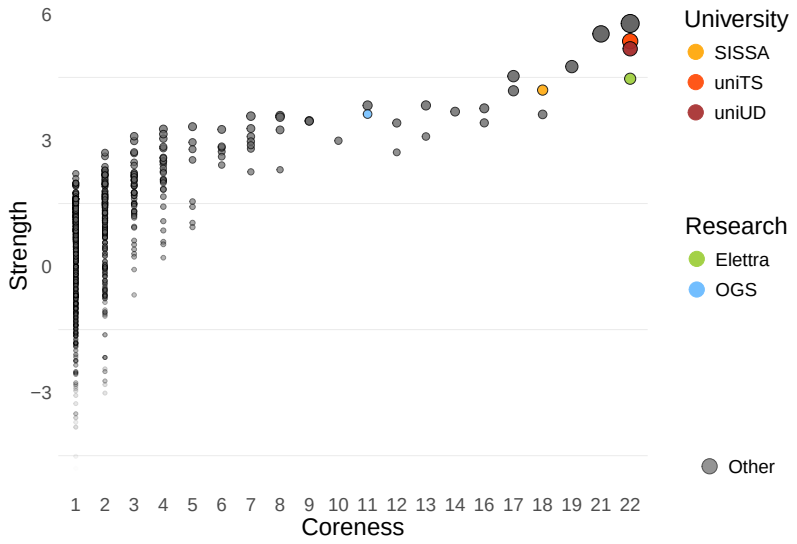
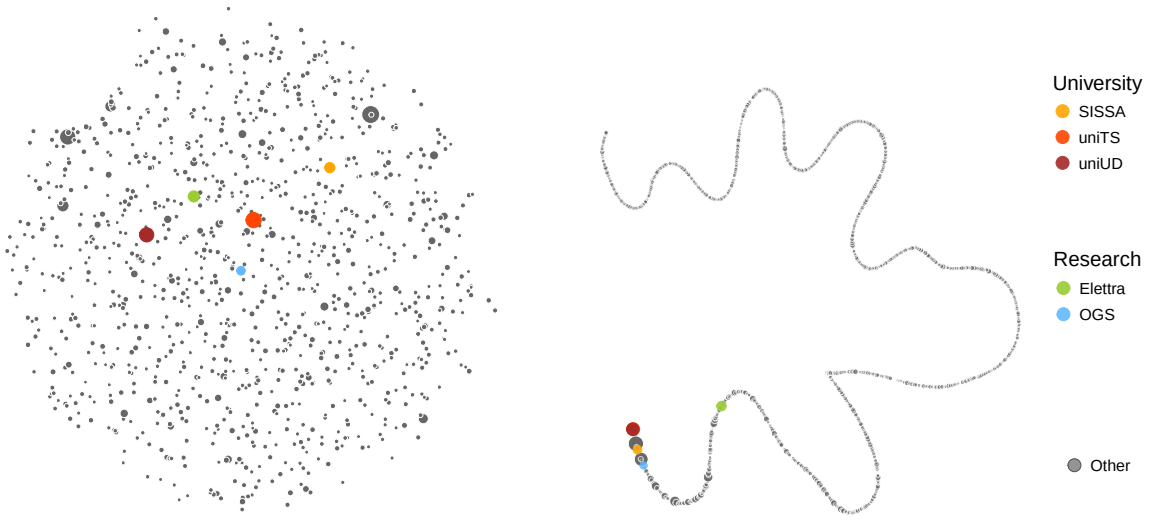


FIGURE 1.2: Visualisation of coreness and strength. Node size represents the strength, while node colour represents universities and public research organisations.

an applied perspective, this lack of evidence for well-defined communities may indicate that worker mobility is not organised around stable groups of employers. Rather than being segmented into distinct labour submarkets, mobility appears to be relatively diffuse, with workers circulating across organisations without being strongly confined to specific clusters. This suggests an integrated labour market in which skill distribution is comparatively uniform and not dominated by clearly separated communities of employers. This relatively diffuse mobility pattern reflects both the ISCO selection (highly skilled professionals with transferable expertise) and the inherent mobility bias of transition-based data. More specialised or less mobile cohorts would likely show stronger skill-based segmentation. The right plot shows the universities very close together in space, along with the PROs, illustrating the similarity of their roles in the network. The size of the points represents node strength, which is closely associated with high coreness (as shown in Figure 1.2), and this combination reveals the centrality of universities and PROs. It indicates that, although they share qualified employees with different organisations, they play the same role as central hubs within their respective neighbourhoods.

Figure 1.4 helps in summarising the results of the K-means clustering. The S_k vs K plot in panel a) shows that there is no clear "elbow" (a point of inflexion on the curve), hence we choose the number of clusters K equal to four since it is the point after which S_k begins to decrease more smoothly. Panel b) shows the UMAP visualisation for data partition into four groups. Although the clustering does not yield sharply separated groups, the spatial layout of the points is clearly hierarchical: nodes in the "head" of the distribution occupy a distinct region of the embedding



(A) Proximity-based embeddings *(B) Role-based embeddings*

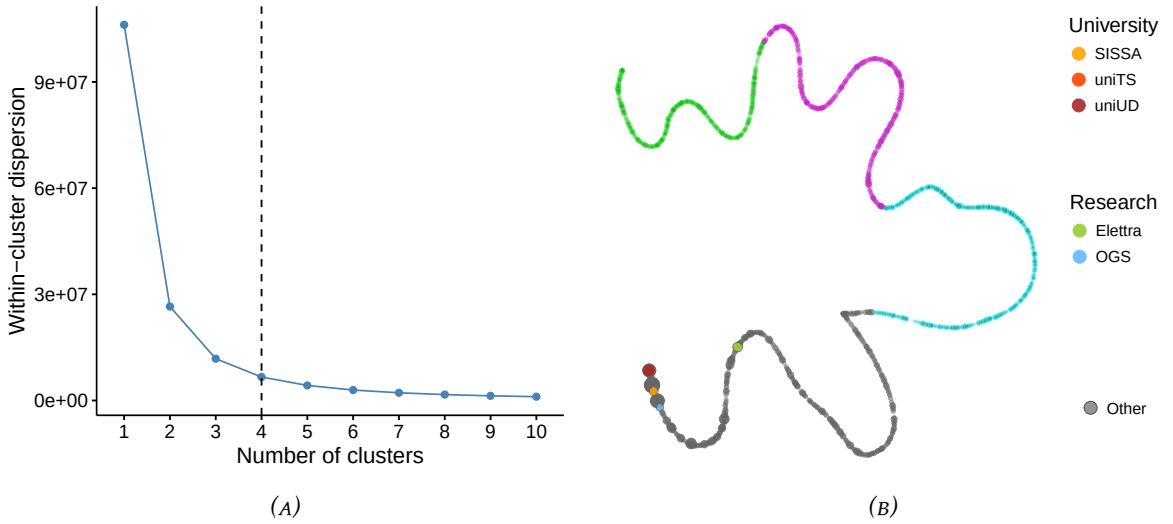
FIGURE 1.3: Two-dimensional UMAP visualisation of node embeddings generated from the LFN using (a) Node2vec and (b) Role2Vec.

space and correspond to a different role with respect to those in the "tail," so we interpret the resulting clusters as reflecting distinct structural positions in the network. The results of the cluster analysis demonstrate that coloured nodes and large-sized nodes cluster together in the role-based embedding space. These findings provide additional evidence of the cohesive roles and influential positions held by universities and PROs.

1.4 Final remarks

Overall, the study highlights the potential of graph RL techniques to analyse complex networks and uncover hidden structures and patterns, which provide new outlooks on economic development and innovation. It also suggests that examining different embedding algorithms tailored to specific tasks would be valuable in addressing different research inquiries.

Results show that universities and leading PROs operating in the FVG region share graduates and/or experienced labour force with different, non-overlapping organisations since they are quite different in terms of structural equivalence. On the other hand, when role-based equivalence comes into play, such differences are reduced. This suggests that their relational role in the overall LFN is similar. Therefore, policymakers may find it interesting that directing efforts to these organisations allows them to reach different other economic actors while following the same structural patterns.



(A) (B)

FIGURE 1.4: *(a) Elbow method for choosing the optimal number of clusters K. (b) Two-dimensional UMAP visualisation of Role2vec embeddings partitioned into four groups.*

It is important to note that the observed patterns of relatively diffuse labour mobility are likely shaped by the specific characteristics of the data used in this analysis. The focus on ISCO major groups 2 and 3–professionals and technicians operating in innovative sectors–means that the sample consists of highly skilled workers whose competencies tend to be transferable across a broad range of organisations and industries. Furthermore, the dataset captures only observed job transitions, inherently selecting for a more mobile subset of the workforce, while workers in stable long-term positions or those with highly specialised, non-transferable skills are underrepresented. It is therefore plausible that applying the same analytical framework to different occupational cohorts–such as manual workers or service sector employees–would reveal more pronounced skill-based clustering and more clearly segmented labour submarkets. Future research should seek to extend this type of analysis to broader occupational groups in order to assess the generalisability of these findings.

Moreover, given that employment contract data is available in the same format across all Italian regions, this analysis can be reproduced in different contexts. Analysing the LFN in FVG allows us to understand labour mobility and knowledge spillovers in a medium-sized region with notable innovation activities. It would be interesting to compare FVG results with those from other regions, such as the larger innovative regions of Lombardia and Emilia Romagna or regions hosting large universities and PROs like Lazio. In future applications, exploring the impact of incorporating node attributes in the RL task could yield valuable insights.

Chapter 2

Spectral clustering of Berlin's housing spatial network

2.1 Introduction

Understanding the determinants of housing prices is essential for both urban planning and policy development. As highlighted by Ryu et al. (2024), accurately measuring the value of a location is a key component of mass appraisals of residential properties, house price index estimation, property taxation, and the evaluation of construction projects. To make informed decisions, investors, but also local governments, rely on accurate models reflecting locational value, thus improving the fairness of property taxation.

Over the years, economists, urban planners, and researchers from various fields have developed a wide array of quantitative models to estimate the market value of housing. The theoretical foundation of these efforts was laid by Rosen (1974), whose seminal contribution formalised how product differentiation in competitive markets leads to equilibrium prices that reflect the value of individual product attributes. Building on this framework, a vast literature has applied hedonic pricing models (HPMs) to study housing markets, environmental amenities, and various aspects of product quality.

A major strength of hedonic models is their ability to decompose a property's price into the contribution of its individual characteristics, offering both flexibility and interpretability. However, this simplicity often comes at the cost of predictive accuracy. Recognising these limitations, scholars have proposed various modifications to improve hedonic models, particularly with the aid of pre-processing techniques, spatial analysis with Geographic Information Systems (GIS), and advanced machine learning algorithms. Wei et al. (2022) provides an extensive review of over 100 studies in this domain. Despite the key developments highlighted in the review, capturing locational effects accurately remains a challenging but crucial task for improving housing price models, with limited contributions in the existing literature addressing this issue.

A promising approach in this regard was proposed by Ryu et al. (2024), who introduced a two-stage framework that incorporates fixed location effects via K-means

clustering. It is important to note that alternative clustering techniques may provide superior performance in capturing spatial variation. In this paper, we extend this approach by leveraging a rich dataset from ImmoScout24–Germany's largest platform for residential and commercial property listings–focusing on the Berlin housing market in 2021 and 2022. Alongside detailed property characteristics, we incorporate geospatial information using two clustering techniques to improve the modelling of locational effects: K-means and spectral clustering. Both methods are applied in their constrained forms to account for the heterogeneity in housing density across neighbourhoods. While both methods provide a more granular and flexible spatial representation compared to administrative boundaries, spectral clustering proves particularly effective in capturing the geographic relationships between neighbouring cells, further enhancing the spatial component of the hedonic model. In this approach, sparse areas are grouped together into larger clusters, while dense areas are allowed to form smaller clusters, including isolated individual cells. To the best of our knowledge, we are the first to apply spectral clustering to estimate locational effects within a hedonic pricing framework.

Improving the estimation of locational value is not only crucial for enhancing the accuracy of property valuation models but also has important implications for public policy. More accurate assessments of locational effects allow property taxation systems to better reflect the actual market value of dwellings, thereby improving taxation fairness. In turn, this reduces the risk that wealthier households end up benefiting from undervaluation, which would otherwise contribute to increasing horizontal inequity and exacerbating income and wealth inequalities within urban areas (Sheffrin et al., 2008; Gilderbloom et al., 2012).

The case of Berlin represents an especially relevant application, given the critical importance of location in shaping property values in metropolitan contexts. However, estimating these effects remains challenging due to Berlin's spatial heterogeneity of urban environments.

The remainder of the paper is structured as follows. Section 2.2 provides a brief review of the relevant literature. Section 2.3 introduces the data and Section 2.4 presents the methodological framework, including the clustering techniques and hedonic models (Sections 2.4.1 and 2.4.2). Section 2.5 reports the empirical findings, comparing the performance of alternative approaches to modelling locational effects. Section 2.6 concludes by discussing the implications of our findings for urban policy and real estate valuation, with particular emphasis on the benefits of incorporating refined spatial indicators into property price models.

2.2 Literature review

2.2.1 Hedonic pricing models

Since Rosen (1974) seminal contribution, the HPM has become the standard framework for estimating the implicit value of non-directly observable characteristics of goods, particularly in the housing market. By decomposing the price of a good into the value of its constituent attributes, this approach allows researchers to disentangle the contribution of structural housing features, neighbourhood characteristics, and location-specific factors to property values, along with other determinants. The main strengths of hedonic models lie in their simplicity and interpretability, particularly through the estimation of the implicit contribution of individual attributes to the overall asset price. However, a persistent limitation of this framework is that many of the factors influencing housing prices–especially those related to location–are difficult to observe or quantify, making them challenging to incorporate into conventional regression models.

Over the past decades, a growing body of research has sought to address these limitations from multiple angles. Scholars have experimented with several enhancements to the basic model. It is also worth noting the important role of the spatial econometrics literature in the development of hedonic pricing models. Seminal contributions such as Dubin (1992), which examined spatial autocorrelation in hedonic house price models, and Anselin (1988) laid the conceptual foundations for modelling spatial dependence in property values. These frameworks were further developed and extended by LeSage and Pace (2009a,b, 2014), providing strong methodological tools and applications across a wide range of economic and real estate contexts, as well as by numerous other researchers. More recent methodological advances include geostatistical and semiparametric spatial models for real estate pricing (e.g., Muto et al. (2023); Schirripa Spagnolo et al. (2024)), which further enhance the modelling of spatial heterogeneity and autocorrelation in property markets, and represent a methodological backdrop for the subsequent extensions put forth by the analysis of locational effects presented below.

2.2.2 Capturing locational effects

While many contributions have significantly improved the predictive power of hedonic models, one area remains relatively underexplored: the modelling of locational effects. As anticipated in the introduction, traditional hedonic models typically capture location effects through linear terms or coarse spatial dummies, which often fail to account for the complex, nonlinear spatial dynamics characterising urban real estate markets.

The literature has provided extensive evidence of how locational factors shape housing prices, particularly through their interaction with local labour markets. Early

contributions, such as Johnes and Hyclak (1999) link housing prices with labour market conditions like wages, unemployment rates, and labour force participation.

Wheaton and Lewis (2002) and Osland and Thorsen (2008) show how urban agglomeration and labour market accessibility shape wage variation. Moretti (2011, 2012, 2013) further develops this perspective, analysing how cities offering better employment opportunities tend to attract skilled labour, thereby increasing housing demand and, consequently, housing prices. This mechanism, in turn, amplifies spatial inequalities across urban areas.

Educational amenities, especially school quality, have been widely studied as influential locational effects in HPM (Haurin and Brasington, 1996; Kane et al., 2006; Feng and Lu, 2013). However, it is important to emphasise that the correlation between school quality and housing prices does not necessarily imply causality. Higher-income families are more likely to live in expensive neighbourhoods and, at the same time, to invest more in their children's education, making it difficult to disentangle the causal relationship between school quality and house prices. Recent literature has also begun to explore the interplay between school choice mechanisms, residential decisions, and housing market dynamics (Reback, 2005; Bayer et al., 2007; Avery and Pathak, 2021).

Beyond education, cultural amenities have increasingly been recognised as important drivers of housing prices by contributing to neighbourhood attractiveness (Sheppard, 2010; Moro et al., 2013; Borgoni et al., 2018).

Environmental factors also represent an important category of locational effects frequently analysed through HPM, particularly in relation to environmental externalities. Studies (Brasington and Hite, 2005; Nicholls, 2019; Miłuch and Kopczewska, 2024) explore how pollution and environmental disamenities impact housing prices, though data limitations often undermine precise estimations.

Moreover, even advanced machine learning approaches often struggle to effectively disentangle locational effects from structural housing characteristics, largely due to their typically lower interpretability (Ryu et al., 2024). As a result, accurately capturing locational effects remains a challenging yet crucial task for enhancing the performance of housing price models.

Indeed, as previously discussed, even standard hedonic models often face inherent limitations in fully incorporating the wide range of relevant locational covariates, primarily due to constraints in data availability. This highlights the persistent need for methodological innovation to improve the treatment of spatial heterogeneity in property valuation models.

Several studies have sought to address this issue by experimenting with alternative definitions of spatial units and neighbourhood structures. Kryvobokov (2013), for example, proposed redefining neighbourhood boundaries using Thiessen polygons constructed around individual observations, combined with a fuzzy clustering approach to be embedded within HPM. His findings showed mixed results–improvements were more evident when using simple OLS estimation–yet the study was one of the

first to emphasise the advantages of moving beyond conventional administrative boundaries.

Expanding on this line of inquiry, Kwon et al. (2017) introduced an innovative methodology for delineating urban areas in Seoul by applying K-means clustering in conjunction with hedonic models. Their results demonstrated clear improvements in price prediction accuracy when compared to models based on traditional administrative divisions. Similarly, Calka (2019) advanced the field by combining clustering techniques (specifically, K-means clustering) with geostatistical methods such as kriging, which allowed her to interpolate and smooth spatial patterns of housing prices while accounting for the underlying spatial dependence in the data. This hybrid approach provided a more realistic and fine-grained representation of spatial price variation, particularly valuable in urban environments characterised by strong heterogeneity in housing markets.

Among the most promising contributions in recent years is the two-stage framework developed by Ryu et al. (2024). Their approach first partitions the housing market into clusters based on proximity by applying K-means clustering to the geographic coordinates, and then uses these clusters as fixed locational effects in a subsequent hedonic regression. The benefits of this method are multiple: it requires only basic geographic coordinates (latitude and longitude), accommodates both nonlinearities and unobserved locational factors without imposing a restrictive functional form, and consistently yields better predictive accuracy compared to traditional hedonic specifications. Crucially, this approach offers a meaningful conceptual separation between structural housing characteristics and pure locational effects.

Beyond these approaches, spatially constrained clustering offers additional tools for improving the modelling of locational effects. The Spatial 'K'luster Analysis by Tree Edge Removal (SKATER) algorithm partitions space by constructing a minimum spanning tree based on spatial contiguity and attribute similarity, and then pruning edges to obtain geographically contiguous and internally homogeneous clusters (Assunção et al., 2006). In other words, the spanning tree encodes the adjacency relationships between spatial units, and this network-based spatial constraint represents a key advantage over techniques such as K-means: it produces clusters that more closely reflect real neighbourhood structures and are therefore better suited to capturing localised price variation. A recent extension, SKATER-reg (Anselin et al., 2023), incorporates regression-based criteria into the clustering process, producing spatial regimes that minimise within-cluster hedonic prediction error. This makes SKATER-reg especially useful for HPM applications in which spatial heterogeneity is unlikely to align with administrative boundaries or standard functional forms.

In this context, Ferligoj and Batagelj (1982) can be seen as one of the early contributions to clustering with relational constraints. Their work formalises the idea that clusters should respect underlying connectivity patterns, which resonates strongly with later spatially constrained clustering methods such as SKATER. However, a crucial limitation for our study is that most spatially informed methods rely on detailed spatial adjacency information to construct the underlying network.

In parallel to these algorithmic approaches, a substantial literature has developed around model-based clustering methods for spatially dependent data, often relying on probabilistic frameworks such as Markov random fields or conditional autoregressive structures (Alfò et al., 2008; Alfó et al., 2009). While model-based spatial clustering offers important theoretical advantages, these methods typically require detailed spatial adjacency matrices, impose distributional assumptions, and can be computationally demanding, especially for large datasets.

In many real estate applications–including the present study–geographic data are only available in aggregated form, preventing the direct use of these methods. This necessitates the development of an alternative clustering strategy that explicitly incorporates the available aggregation structure when constructing the spatial network, a challenge that is common across empirical housing market research.

Building on these observations, the objective of the present study is to explore alternative clustering techniques that may offer superior performance in capturing spatial variation in housing prices, thereby addressing some of the unresolved challenges in the empirical modelling of real estate markets.

2.3 Motivating example

Berlin offers a compelling setting to illustrate our approach, as its housing market combines strong spatial heterogeneity with decades of institutional restructuring, privatisation, and uneven neighbourhood development. These dynamics have produced a highly segmented urban landscape in which price formation depends not only on property attributes but also on localised market pressures, investor activity, and differentiated trajectories of gentrification and redevelopment. To capture these fine-grained spatial patterns, we focus on a recent and internally coherent period and draw on high-resolution microdata, as detailed below.

The analysis is restricted to the years 2021 and 2022, which represent the most recent period available in the dataset and allow us to analyse contemporary market conditions. In line with Arlia (2024), we exclude 2020 to avoid distortions in market dynamics associated with the COVID-19 pandemic. Focusing on this two-year window minimises disruptions such as listing suspensions and irregular transaction patterns, resulting in a more consistent and homogeneous dataset. We also refrain from extending the temporal window further back, as doing so would reduce temporal uniformity and introduce heterogeneity across market regimes; instead, we prioritise a recent and internally consistent sample.

We use data from ImmoScout24 (Schaff and Thiel, 2023), the largest real estate advertisement platform in Germany. The dataset provides detailed information on property characteristics–such as living area, available amenities, and construction year–together with spatial referencing at the 1 km^2 grid level, which represents the finest spatial resolution available. Data covers both rentals and sales, for both houses and flats. Since the markets for flats and houses have structural differences and

therefore need to be analysed separately, this study focuses exclusively on flats, which offer greater homogeneity, improving the accuracy of hedonic estimates. Furthermore, we focus on sales prices rather than rents for the Berlin market of 2021-2022. Although a substantial body of research finds sales prices generally more volatile than rents (Gallin, 2008; Hilber and Mense, 2021; Miles, 2025), our data suggest that, in Berlin during 2021–2022, sales prices exhibited relatively smoother dynamics. This likely reflects both the post-pandemic acceleration in rent inflation in the euro area (Arioli et al., 2023) and institutional features of the Berlin rental market, which has historically been characterised by rent controls and contract rigidities, and in these years was experiencing the gradual unwinding of policies such as the Mietendeckel, historically moderating short-run adjustments in rents. Consequently, sales prices provide a suitable measure for capturing structural pricing patterns in Berlin over this period.

The sample is restricted to those units for which complete information on the main observed characteristics is available and meet pre-defined criteria. After applying these filters, the final sample consists of 90,849 observations, distributed over 611 grid cells of 1 km^2 (Figure 2.1).

To refine the dataset, several variables are recategorised. The detailed heating system types are grouped into broader categories by combining similar heating types, reflecting modern or eco-friendly, not eco-friendly, and unspecified heating.

The number of rooms variable is recategorised into three groups: small (up to 2 rooms), medium (more than 2 but up to 3 rooms), and large (more than 3 rooms).

The missing information in the parking variable has been re-coded as "No" and a new variable "Parking Coverage" has been created, measuring the percentage of available parking spaces relative to the total number of flats in the same grid.

To incorporate neighbourhood quality indicators into the analysis, we utilised the RWI-GEO-GRID dataset. This dataset offers socio-economic data for Germany, structured at the same $1\,\text{km}^2$ resolution, enabling consistent merging with the ImmoScout24 data. It is useful to include socio-spatial variables in the model to account for key determinants of property values and to capture observed spatial heterogeneity across neighbourhoods. Specifically, household density and commercial units reflect land-use intensity and the mix of residential and commercial activity, which influence local amenities and externalities relevant for prices. Likewise, car density proxies mobility patterns and transport dependency, foreign household share serves as an indicator of socio-demographic composition, and children per household signal the presence of families, a feature typically associated with school quality, safety, and overall neighbourhood livability. The unemployment rate conveys local labour-market conditions and economic vitality, directly affecting neighbourhood attractiveness, while housing type reflects the scale and typology of the housing stock, influencing both supply characteristics and buyer preferences. While we acknowledge that buyers do not explicitly take these variables into account, they provide information on underlying neighbourhood attributes that indirectly affect housing prices. As such, they help capture both observable and latent spatial effects and reduce the risk of omitted-variable bias.

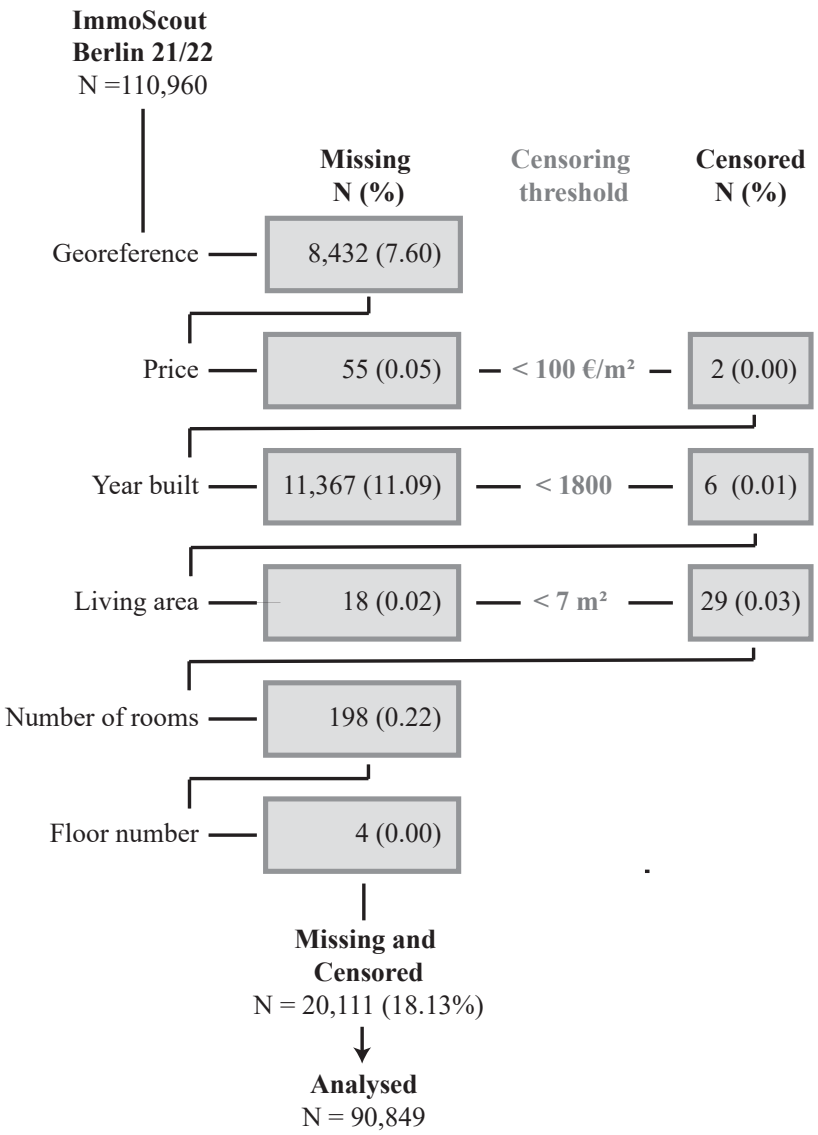


FIGURE 2.1: Flow chart summarising data cleaning and selection steps used to construct the final sample.

District-level information was integrated using the Berlin districts shapefile (Figure 2.2). All spatial analyses are conducted at the $1\,km^2$ grid level, each grid corresponding to a square on the map. To support the methodological discussion and provide a unified overview of the dataset, Table 2.1 presents all variables employed in the analysis, along with their definitions and corresponding data sources.

2.3.1 Descriptive statistics

Information on various residential properties attributes is summarised in Table 2.2 in terms of central tendency and distribution for numerical variables, and frequency

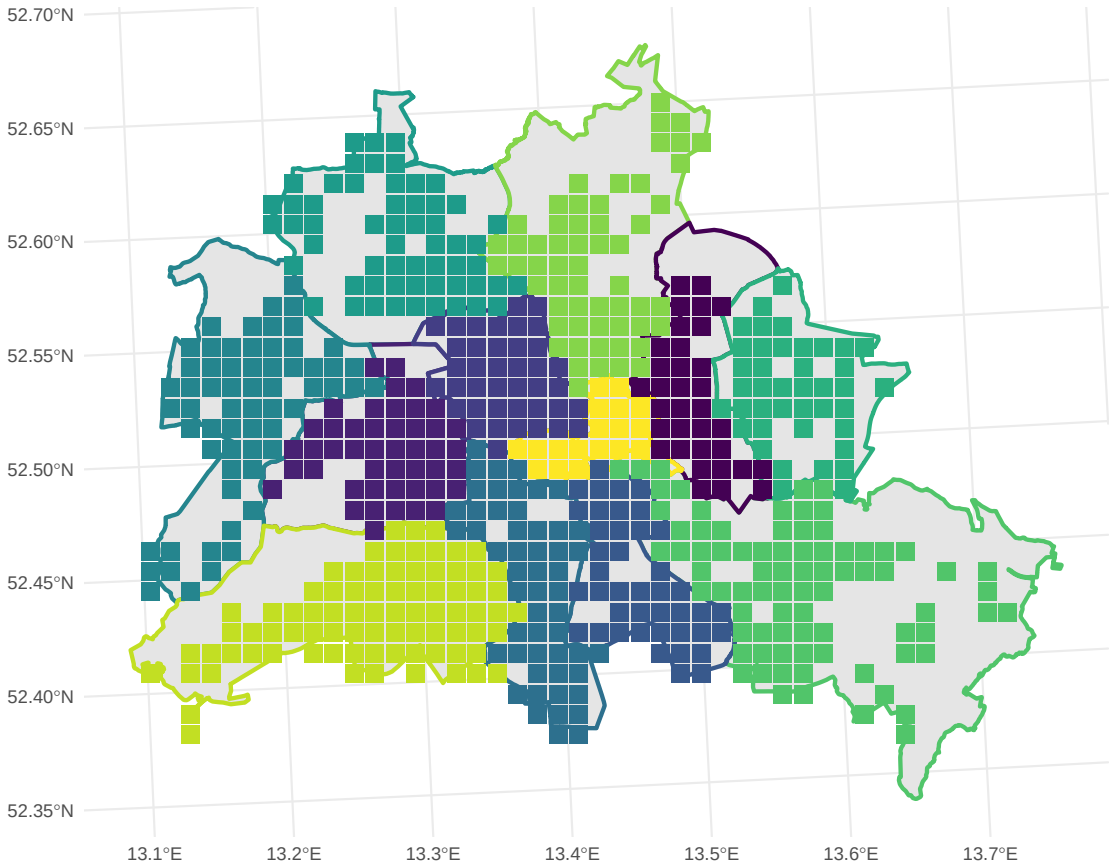


FIGURE 2.2: Berlin districts map and 1 km^2 grid level map, with cells coloured according to district membership.

distribution for categorical variables. The average offer posted selling price of properties is 550,483.32 €, with a substantial variability (standard deviation of 549,764.64 €) and median price of 385,000.00 €, suggesting a positive skew in the distribution (i.e., more properties priced below the mean). The median construction year is 1956, with many properties built in the mid-20th century. The mean living area is $82.76\,m^2$, with considerable variation (standard deviation of $46.55\,m^2$). The majority of properties (60.8%) do not have an elevator. Approximately half of the properties (49.0%) have 2 or fewer rooms, which also comprises half rooms (a room with a size between 6 and 10 m^2). Properties with more than 2 but less than or equal to 3 rooms constitute 30.1%, while 20.9% of properties have more than 3 rooms. A large proportion of properties (73.3%) have either a balcony or a garden, and 25.3% have modern or eco-friendly heating types.

Neighbourhood-level characteristics from the RWI-GEO-GRID dataset are summarised in Table 2.3. The dataset reveals significant variation in the number of private households per grid cell, with a mean of 3,129.90 and a median of 2,175.00, reflecting a high standard deviation of 2,769.40. Similarly, the number of commercial buildings

varies considerably, with an average of 415.05 and a median of 226.00.

The proportion of foreign households averages 16.0%. The children per household ratio remains relatively consistent, averaging 0.25, with little variation across the grid cells. The unemployment rate displays moderate variability, averaging 7.5% with a

Variable	Definition	Source
Property characteristics (ImmoScout24)		
Price	Purchasing price in EUR	ImmoScout24
Living Area	Living space in square meters	ImmoScout24
Construction Year	Year the flat was built	ImmoScout24
Elevator	Presence of elevator	ImmoScout24
Number of Rooms	Recategorized in: ≤ 2, > 2 and ≤ 3, > 3 rooms	ImmoScout24
Balcony or Garden	Presence of balcony or garden	ImmoScout24
Heating Type	Recategorised in: eco-friendly, not eco-friendly, other	ImmoScout24
Parking	Missing values recoded as "No"	ImmoScout24
Parking Coverage	Share of flats with parking within the same 1 km^2 grid cell	Constructed variable
Neighbourhood characteristics (RWI-GEO-GRID)		
Households	Number of private households	RWI-GEO-GRID
Commercial Buildings	Number of commercial buildings	RWI-GEO-GRID
Car Density	Cars per household	RWI-GEO-GRID
Foreign Households	Share of households with foreign head of household	RWI-GEO-GRID
Children	Number of children per household	RWI-GEO-GRID
Unemployment Rate	Share of unemployed residents	RWI-GEO-GRID
Housing Type	Recategorised in: 3–5 Family Houses, 6–9 Family Houses, 0–19 Households, > 20 Households	RWI-GEO-GRID
Spatial reference		
Grid Cell ID	Spatial unit at the 1 km^2 resolution (finest available level)	ImmoScout24 & RWI-GEO-GRID
District ID	Berlin administrative district	Berlin Shapefile

Table 2.1: Overview of variables, definitions, and data sources used in the analysis.

standard deviation of 3.6%. The parking coverage variable indicates that, on average, 22.0% of flats within a 1 km^2 area have parking available. On average, car density is 0.64 cars per household.

In terms of housing type distribution, the majority of grid cells are characterised by larger residential buildings. Specifically, half of the cells predominantly feature buildings with more than 10 households.

2.4 Methodology

2.4.1 Clustering

The first step in this work is to subdivide the location plan by taking into account the spatial distribution of flats for sale. The distribution of these properties is typically irregular, influenced by city topography, transport infrastructure, and other socio-economic variables.

A simplistic subdivision approach using the 1 km^2 grid is unsuitable. Although it provides a uniform structure, it does not take into account variations in the density of housing samples within the city. Cells with few flat listings lack representative characteristics.

		Mean	SD	Median
Price		550,508.32	549,768.64	385,000.00
Construction Year		1953	46	1956
Living Area		82.76	46.55	70.00
	Level		N	%
Elevator	No		55,244	60.8
	Yes		35,605	39.2
N Rooms	≤ 2		44,561	49.0
	> 2 & ≤ 3		27,340	30.1
	> 3		18,948	20.9
Balcon or Garden	No		24,242	26.7
	Yes		66,607	73.3
Heating Type	Modern or Eco-friendly		22,957	25.3
	Not Eco-friendly		54,881	60.4
	Other		13,005	14.3

TABLE 2.2: Summary of characteristics for flats listed for sale in Berlin, 2021–2022 (N = 90,849).

Using administrative districts as a framework for capturing location effects also proves inadequate. Districts are not designed for analytical precision, which leads to grids that are not adequately segmented. As a result, they fail to reflect granular variations in neighbourhood characteristics, which are essential for constructing accurate price models.

One solution could be to divide the 1 km^2 cells into disjoint sets based on the similarity of geographic location.
Ryu et al. (2024) have proposed treating latitude and longitude coordinates as numerical features and using K-means clustering, which allows control over the level of spatial approximation as it involves choosing the number of clusters. In our study, we apply a minimum cluster size-constrained version of the K-means clustering following Ganganath et al. (2014), implemented in R using the base `kmeans` function with a custom wrapper that enforces a minimum cluster size. The procedure first runs standard K-means on the geographic coordinates to obtain an initial partition. Clusters with fewer than a specified minimum number of observations are then identified and suppressed, and the corresponding cells are reassigned to the nearest remaining cluster centre. This ensures that no cluster falls below the minimum size threshold, which in our case is set to 20 observations. Higher-density regions result in a greater number of clusters, while the constraint ensures that clusters with very few observations are avoided. This guarantees consistent estimation quality across areas with varying densities.

An alternative approach constructs a spatial network based on geographical

		Mean	SD	Median
Parking Coverage		0.22	0.20	0.16
Household		3,129.90	2,769.40	2,175.00
Commercial		415.05	557.20	226.00
Car Density		0.64	0.19	0.65
Foreign Household		15.98	8.01	13.96
Children per Household		0.25	0.05	0.25
Unemployment Rate		7.46	3.64	7.72
	Level		N	%
House type	1-2 Family Houses		197	32.2
	3-5 Family Houses		14	2.3
	6-9 Family Houses		96	15.7
	10-19 Households		143	23.4
	> 20 Households		161	26.4

TABLE 2.3: Summary of characteristics for 1 km^2 Berlin raster cells in 2017 (N = 611).

distances (Assunção et al., 2006) and applies spectral clustering (von Luxburg, 2007) in its constrained version. We introduce a novel method for spatial network construction designed specifically for contexts in which only aggregated geographical information is available. In our proposed framework, each cell in the 1 km^2 grid is connected by an edge to its adjacent cells–defined as those whose coordinates differ by only one unit–thereby establishing a simple and consistent adjacency structure. This adjacency rule ensures that each cell is connected only to its immediate neighbours, forming a regular lattice-like network that reflects the underlying spatial layout of the study area. To account for the heterogeneity inherent in aggregated data, we assign edge weights inversely proportional to the total number of observations in the two connected cells. This weighting scheme makes sparsely populated cells more likely to cluster together, while denser cells are able to form more refined, standalone groups. In this way, our approach explicitly leverages the aggregation structure of the data and adjusts for differences in population density, allowing the constructed spatial network to remain informative even when fine-grained geographic detail is unavailable.

Spectral clustering proceeds by computing the graph Laplacian and extracting its leading eigenvectors. We then apply the constrained K-means clustering to these eigenvectors. The number of clusters is set to 303, implying an average size of approximately 300 flats per cluster, with a minimum cluster size of 20. This lower bound is imposed to ensure statistical stability in the subsequent hedonic regressions and to prevent the formation of sparse clusters, which would undermine the reliability of coefficient estimates. The choice of K is guided by the empirical distribution of listings and reflects a balance between two competing objectives: achieving sufficient within-cluster homogeneity in locational effects while retaining enough observations per cluster to support stable inference.

Although one motivation for clustering is dimension reduction, the degree of reduction is necessarily constrained in settings with highly uneven spatial data density. In our case, specifying a small K or a loose minimum-size constraint would have produced overly large clusters that mask spatial heterogeneity–precisely the variation the method aims to uncover. The selected configuration, therefore, represents a practical compromise between granularity and statistical adequacy.

However, to assess the strength of the clustering specification, we also estimated models using two alternative configurations corresponding to approximately 200 and 400 flats per cluster. The results–available in the appendix–confirm that our main findings remain qualitatively unchanged under our alternative spatial subdivision. However, incorporating locational attributes derived from Spectral Clustering with $K = 303$ yields improved computational efficiency by operating on a much smaller number of nodes (611 cells versus 90,849 apartments) and quantitatively superior predictive performance.

2.4.2 Hedonic pricing models

In the second stage, we estimate a set of HPMs that differ in how they capture locational effects. Specifically, we compare three model specifications based on alternative locational indicators, along with a baseline model that omits locational controls. The general model specification takes the following form:

$$\ln P_i = \beta_0 + \sum \beta_s X_{s,i} + \sum \beta_n X_{n,i} + \sum \beta_l X_{l,i} + \epsilon_i,$$

where P_i is the price of flat i (the dependent variable), and X represents the independent variables, which are grouped into structural (X_s), neighbourhood (X_n), and locational (X_l) attributes. While this specification assumes a linear relationship between the logarithm of prices and the covariates, this formulation is standard in HPM and provides a transparent and interpretable baseline for assessing the relative contribution of alternative locational indicators.

Structural attributes include the logarithm of the living area, the number of rooms, the availability of an elevator, the presence of a balcony or garden, the type of heating system, and the year of construction. Neighbourhood attributes comprise parking coverage, car density, the number of commercial buildings, the number of households, and the housing type, as well as the number of foreign households, children per household, and the unemployment rate within each grid cell. The locational attribute is a factor with levels representing administrative or cluster assignments. The first hedonic model incorporates administrative districts as location indicators. The second employs clusters generated through constrained K-means with $K = 303$, while the third model, representing our proposal, utilises spectral clustering with $K = 303$ as the locational indicator. Spectral clustering is particularly advantageous due to its capacity to effectively capture spatial dependencies by considering both geographic proximity and variations in housing density, thereby providing a more accurate spatial representation within the model.

Additionally, as anticipated, we extend the comparison to include a baseline model that excludes locational information entirely, providing a reference point to assess the impact of incorporating spatial attributes on predictive performance.

To evaluate the model's prediction accuracy, we used three standard metrics: Root Mean Squared Error (RMSE), which measures the average error between observed and predicted values, with lower RMSE indicating a better model fit; R-squared (R^2), which represents the proportion of variance in the dependent variable explained by the independent variables, with higher R^2 values indicating improved explanatory power; and Mean Absolute Error (MAE), which assesses the average magnitude of prediction errors, regardless of their direction. Prediction performance is assessed using five-fold cross-validation.

Beyond assessing predictive performance, we investigate which variables have the greatest impact on predicting flat sale prices. For this purpose, we compute standardised feature importance values within the best-performing hedonic model. The importance of each variable X_q is calculated as the product of its estimated

FIGURE 2.3: Flats for sale density in $1\,km^2$ *cells in Berlin in 2021-2022.*

coefficient β_q and the standard deviation of X_q; the value is rescaled to a 0-100 range for ease of interpretation (Gevrey et al., 2003; Kuhn and Max, 2008).

2.5 Application: the Berlin's housing spatial network

2.5.1 Clustering

Figure 2.3 illustrates the spatial distribution of flats for sale across Berlin, with each square representing a cell in the $1\,km^2$ grid. Darker colours indicate cells with a higher concentration of listings. This grid division results in a highly segmented location plan. The map highlights the highly uneven spatial distribution of housing supply, with dense cells located primarily in central and well-connected areas, while peripheral regions exhibit significantly lower densities. Employing this partitioning of the city in the model would lead to inconsistent estimates due to insufficient data.

To address this issue, we apply two spatial clustering approaches that aggregate neighbouring cells based on both proximity and listing density, ensuring sufficient observations within each spatial unit for stable estimation. The results of the clustering procedures are presented in Figure 2.4, where colored blocks represent different clusters. When incorporating a minimum cluster size constraint, the resulting number of clusters in constrained K-means is $K = 259$ for an initial K of 303, while the spectral clustering results in 249 constrained clusters. The slight reduction in the

(A) (B)

FIGURE 2.4: Spatial clustering results using (a) constrained K-means on cell coordinates with $K = 303$, (b) constrained spectral clustering on weighted spatial network with $K = 303$.

number of clusters relative to the initial K reflects the merging of low-density cells to meet the minimum observation threshold, ensuring that each cluster contains a sufficient number of listings for reliable analysis. Notably, both clustering methods produce spatial units that better align with the underlying distribution of flats for sale compared to administrative districts or the raw grid. Spectral clustering, in particular, demonstrates an enhanced ability to respect geographic contiguity and local density variations, producing spatial clusters that more accurately capture neighbourhood effects relevant for price modelling.

2.5.2 Hedonic pricing models

The four estimated models differ exclusively in how they incorporate locational attributes, ensuring that any observed improvement in predictive accuracy is directly attributable to the model's ability to capture the complexity of neighbourhood effects. Any differences underscore the crucial role of accurately modelling spatial influences in the predictive performance of HPMs.

Table 2.4 presents the predictive performance of the models using three metrics (RMSE, R^2, and MAE), showing the impact of the different locational representations on predictive accuracy. Our proposed specification, which employs constrained spectral clustering as the locational indicator, demonstrates the strongest performance among all alternatives. It achieves the lowest RMSE (0.25), indicating the smallest average prediction error, and the highest R^2 value (0.85), meaning that approximately 85% of the variance in sale prices is explained by the model. The MAE of 0.19 further reflects minimal average deviation between predicted and observed prices.

To facilitate comparison across models, Table 2.4 also reports percentage changes relative to the baseline model without locational controls. Improvements in RMSE and MAE are computed as percentage reductions in error:

$$\Delta\text{RMSE}_k = \frac{\text{RMSE}_{\text{baseline}} - \text{RMSE}_k}{\text{RMSE}_{\text{baseline}}} \times 100, \qquad \Delta\text{MAE}_k = \frac{\text{MAE}_{\text{baseline}} - \text{MAE}_k}{\text{MAE}_{\text{baseline}}} \times 100,$$

while the percentage change in explanatory power is computed as:

$$\Delta R^2_k = \frac{R^2_k - R^2_{\text{baseline}}}{R^2_{\text{baseline}}} \times 100.$$

Under these measures, constrained spectral clustering yields the largest gains: approximately a 19% improvement in RMSE, a 21% improvement in MAE, and a 9% improvement in R^2 relative to the baseline. Hence, the model without any locational controls performs substantially worse, confirming that omitting spatial information leads to notable predictive loss. Incorporating administrative districts improves performance (RMSE = 0.27, $R^2 = 0.83$), outperforming the K-means clustering approach, but still less effectively than the spectral clustering specification.

These findings underscore the advantages of constrained spectral clustering in flexibly capturing spatial structure and heterogeneity relevant for property valuation–advantages that neither administrative boundaries nor density-agnostic clustering methods fully exploit. Importantly, the results remain consistent when evaluated on an external test set, confirming the strength of the proposed approach.

In terms of variable significance, nearly all structural and neighbourhood attributes included in the model are statistically significant at the 1% level ($p < 0.001$), except for the "Children per Household" variable, which shows weaker significance ($p < 0.005$).

Locational Attribute	Performance Metrics			Improvement vs. Baseline (%)		
	RMSE	R^2	MAE	ΔRMSE	ΔR^2	ΔMAE
None (Baseline)	0.31	0.78	0.24	–	–	–
Administrative Districts	0.27	0.83	0.21	+12.9%	+6.4%	+12.5%
K-Means Clustering	0.29	0.81	0.22	+6.5%	+3.8%	+8.3%
Spectral Clustering	0.25	0.85	0.19	+19.4%	+9.0%	+20.8%

TABLE 2.4: Predictive performance of hedonic models using different locational attributes, along with percentage improvements relative to the baseline model without locational controls.

Table 2.5 reports the estimated coefficients for the model employing spectral clustering, along with standard errors and significance levels. [1]

Consistent with theoretical expectations, larger living spaces, more rooms, the presence of an elevator, a balcony or garden, and modern heating systems are all positively associated with higher sale prices. Neighbourhood characteristics such as higher parking coverage and lower unemployment rates also correspond to higher property values. Interestingly, greater car density is associated with a negative price effect, suggesting that traffic congestion or limited accessibility detracts from perceived property value.

Table 2.6 reports the standardised variable-importance values, which quantify each predictor's relative contribution to the model's predictive performance. The importance scores are rescaled so that the most influential variable–the logarithm of living area–takes a value of 100, with all other variables expressed proportionally. Because the scores are computed from standardised predictors, a variable with a

[1] We have also computed the location indicators for the other estimated models–including the standard hedonic model without locational attributes, the model with administrative location controls, and the constrained K-means model. These results are available upon request.

TABLE 2.5: Hedonic price model estimates.

	Estimate	Std. Error	Pr(> |t|)
(Intercept)	5.48	6.06×10^{-2}	$< 10^{-10}$
$\log$(Living Area)	1.07	3.01×10^{-3}	$< 10^{-10}$
N Rooms > 2 & ≤ 3	1.24×10^{-2}	2.45×10^{-3}	4.15×10^{-7}
N Rooms > 3	5.25×10^{-2}	3.56×10^{-3}	$< 10^{-10}$
Elevator	1.20×10^{-1}	2.22×10^{-3}	$< 10^{-10}$
Balcony or Garden	4.03×10^{-2}	2.12×10^{-3}	$< 10^{-10}$
Not eco-friendly heating	-1.03×10^{-1}	2.25×10^{-3}	$< 10^{-10}$
Heating type not specified	-9.83×10^{-2}	2.99×10^{-3}	$< 10^{-10}$
Construction Year	1.84×10^{-3}	2.48×10^{-5}	$< 10^{-10}$
Parking Coverage	1.29×10^{-1}	8.36×10^{-3}	$< 10^{-10}$
Households	-1.18×10^{-5}	7.23×10^{-7}	$< 10^{-10}$
Commercial Buildings	8.63×10^{-5}	3.24×10^{-6}	$< 10^{-10}$
Car Density	-8.53×10^{-1}	2.38×10^{-2}	$< 10^{-10}$
Foreign Households	-5.59×10^{-3}	2.76×10^{-4}	$< 10^{-10}$
Children per Household	1.47×10^{-1}	5.20×10^{-2}	4.74×10^{-3}
Unemployment Rate	-2.47×10^{-2}	8.63×10^{-4}	$< 10^{-10}$
3–5 Family Houses	-1.89×10^{-1}	1.73×10^{-2}	$< 10^{-10}$
6–9 Family Houses	-1.63×10^{-1}	7.52×10^{-3}	$< 10^{-10}$
10–19 Households	-7.89×10^{-2}	6.91×10^{-3}	$< 10^{-10}$
$\geq$ 20 Households	-1.03×10^{-1}	8.01×10^{-3}	$< 10^{-10}$

small estimated coefficient may still receive a high importance ranking if it exhibits substantial variability in the data. This explains why construction year, despite having a numerically small estimated coefficient, ranks among the most influential predictors: its wide dispersion across observations leads to a substantial contribution to predictive accuracy.

The ranking indicates that structural attributes, particularly living area, construction year, and presence of an elevator, play the dominant role in explaining price variation. Among neighbourhood characteristics, car density, heating system type, and unemployment rate emerge as significant contributors to price prediction, underscoring the role of both individual house characteristics and local socio-economic context in shaping property values.

2.6 Final Remarks

This study focuses on the locational effects in HPMs and compares the predictive accuracy of different model specifications. To this end, we use a rich dataset from

	Variable Importance
Living Area	100.00
Construction Year	20.88
Elevator	15.21
Not Eco-friendly Heating	12.87
Car Density	10.09
Not specified Heating Type	9.24
Unemployment Rate	8.04
Commercial Buildings	7.49
6-9 Family Houses	6.08
Foreign Households	5.70
Balcon or Garden	5.33
Households	4.58
Parking Coverage	4.32
N Rooms > 3	4.14
> 20 Households	3.61
10-19 Households	3.20
3-5 Family Houses	3.06
N Rooms > 2 & <= 3	1.41
Children per Household	0.78

Table 2.6: Standardised variable-importance values for all predictors in the hedonic model with spectral clustering.

ImmoScout24–Germany's largest platform for residential and commercial property listings–focusing on the Berlin housing market in 2021 and 2022. We estimate four HPMs: a traditional specification without locational effects, a standard model including locational controls based on administrative boundaries, one leveraging constrained K-means clustering, and another employing constrained spectral clustering.

The key difference across models lies in how they incorporate locational attributes, ensuring that any improvement in predictive accuracy can be directly attributed to the model's ability to capture the complexity of neighbourhood effects.

Unlike the two more conventional specifications, the clustering-based models are specifically designed to account for the non-uniform distribution of listings and the heterogeneity in housing density across neighbourhoods. While both clustering approaches offer a more flexible and granular spatial representation than administrative boundaries, spectral clustering proves particularly effective in capturing the geographic relationships between adjacent spatial units. In addition, it implicitly incorporates information on cell density, further enhancing the model's ability to account for the spatial distribution of flats for sale.

To the best of our knowledge, this is the first study integrating a spectral clustering algorithm within a standard hedonic pricing framework to estimate locational effects, an approach that improves predictive performance and enhances the accuracy of valuation estimates.

We assess the predictive performance of the models using three widely adopted metrics: RMSE, R^2, and MAE. According to the results, the model incorporating spectral clustering consistently outperforms the alternatives across all three metrics. It is worth noting that, although the other specifications perform worse than the spectral clustering model, they still outperform the baseline hedonic model without any locational controls. Moreover, the model using standard K-means clustering performs similarly to the conventional specification based on administrative boundaries.

In addition to its superior predictive accuracy, the spectral clustering model (as well as the K-means clustering model) offers another key advantage: it requires fewer data inputs than conventional specifications. As shown in the paper, the clustering approaches rely solely on georeferenced information (latitude and longitude), whereas traditional models require access to a much broader range of contextual data.

Berlin represents a particularly relevant case study given the critical role of location in shaping housing prices in metropolitan contexts. To better contextualise our empirical analysis, we also explore Berlin's institutional and regulatory framework, which is characterised by a complex layering of regulatory regimes, harmonisation efforts, and strong public sector involvement in the housing market.

All in all, improving the estimation of locational value is not only crucial for advancing property valuation techniques but also carries significant policy implications. More accurate assessments of locational effects allow property taxation systems to better align assessed values with actual market prices, thereby improving fairness in taxation. In turn, this reduces the likelihood that wealthier households benefit from undervaluation, a phenomenon that would otherwise exacerbate

horizontal inequities and contribute to wealth and income disparities in urban areas (Sheffrin et al., 2008; Gilderbloom et al., 2012).

Additionally, this approach allows for a clearer separation between locational and structural effects within the hedonic model, a feature that holds particular value for urban studies researchers. However, while the two effects remain distinct, clustering-based approaches–both K-means and spectral–do not provide further insights into the specific composition of locational effects.

Naturally, this study has limitations. Although we test the stability of our approach by considering alternative clustering configurations and by replicating the analysis on Hamburg (see Appendices A and B for further details), the results are based on only two cities and a limited time period (2021–2022). Future research should extend the analysis to additional locations with different housing market characteristics, longer time spans, and rental data as well as transaction prices. Applying spectral clustering to estimate locational factors in HPMs across diverse metropolitan housing markets would help assess the generalizability of the model and test its improved predictive accuracy across different contexts and modelling strategies.

Part II

Group formation driven by leader-influence and homophily

Introduction

In social communities, individuals tend to form close relationships with other subjects who share similar traits or attract their interests, resulting in the formation of groups. These groups are typically characterised by denser or more intense connections among their members than between different groups. This structural pattern is commonly referred to as community structure (Newman and Girvan, 2004) and can emerge through different social attachment mechanisms, possibly occurring concurrently or as alternative processes. Triadic closure is the most studied of these mechanisms and represents a natural tendency of real social networks to create group connections (Bianconi et al., 2014), involving trust as a potential aggregation driver. Other mechanisms can give rise to community structures by shaping how new ties are formed, leading to the emergence of cohesive subgroups.

One mechanism is the attraction exerted by influential actors or network leaders, a process determined by the heterogeneity of degree among individuals (Perc, 2014). Highly connected individuals may attract others, promoting the growth of subgroups centred around them. Within these subgroups, leader nodes assume critical roles in the diffusion of information, ideas, and innovations. Consequently, identifying the most influential members within communities has become a key research focus in network science. Another complementary (or even alternative) aggregation mechanism is the propensity for individuals with similar interests or characteristics to bond, known as homophily. Homophily on observed attributes plays a relevant role in shaping the structure and dynamics of networks, contributing to the formation of communities (McPherson et al., 2001).

The mentioned mechanisms do not exist in isolation. Rather, group formation in real social networks could emerge from a complex interplay between structural mechanisms, including triadic closure, preferential attachment associated with degree heterogeneity, and homophily on node attributes (Bachmann et al., 2025). Few contributions propose unified and interpretable frameworks that integrate them jointly—particularly in empirical settings characterised by sparsity, degree heterogeneity, and the availability of rich node attributes. Most studies, focusing on the definition of communities as dense groups, mainly explore the interaction between triadic closure and homophily (Asikainen et al., 2020; Peixoto, 2022; Mosleh et al., 2025). However, growing empirical and theoretical evidence suggests that degree heterogeneity–through the emergence of hubs and bridging actors–plays a central role in shaping communities, interacting with homophily in ways that can either reinforce or blur community boundaries (Bachmann et al., 2025; Fuhse and Gondal, 2024; Kozitsin et al., 2023).

Recent studies on leader-based community detection have demonstrated that, when considering the similarity of attributes together with the presence of leaders, the results tend to be more satisfactory, mainly because the attractiveness of leader nodes may depend on the characteristics of neighbouring nodes (Helal et al., 2016; Lu, 2021; Hu and Wang, 2024). Building on the literature on leader-based community detection, and moving beyond a definition of communities based on density and triadic closure, we argue that leader-driven attraction and homophily jointly shape a structure in which multiple dense, leader-centred cores coexist. Examples of networks exhibiting this pattern include citation networks, where disciplinary proximity generates subgroups led by prominent scholars. For instance, Newman (2001) noted that scientific collaboration and citation networks are governed by star scientists, while also exhibiting high levels of triadic closure.

More generally, scientific collaboration networks provide a key empirical setting for investigating complex group formation mechanisms within the so-called science of science field (Fortunato et al., 2018; Liu et al., 2023). In particular, collaboration networks based on funded research projects exhibit structures that suggest the joint effect of leader-driven attraction and homophily: collaborative ties tend to form around thematic priorities and leading organisations that act as coordination hubs (Morea et al., 2024). The identification of such key players–companies, academic institutions, or research organisations–is crucial for other organisations seeking to engage with them in future projects or for policy makers to evaluate the effectiveness of the funded research in a given region or around a specific topic (Morea et al., 2024).

Motivated by an original dataset on collaborative research, this work extends the analysis of community formation in leader-influenced networks to settings where node-level attributes are available. Building on a density-based community detection (DeCoDe) framework (Menardi and De Stefano, 2022), we integrate structural and attributive information to identify cohesive communities shaped by both connectivity and homophily.

Relevant literature

Community detection in social networks has been extensively studied, with numerous methods developed to address this complex problem (Fortunato and Hric, 2016; Rosvall et al., 2019). Most approaches aim to identify groups of nodes that are more densely connected internally than externally; however, they differ substantially in how density is modelled, estimated, or operationalised. In addition, many assume that all nodes in the network hold equal influence, overlooking the presence of leaders or neglecting the availability of attributive information. Recognising the importance of leader influence and homophily mechanisms, few studies have explored their role in defining the community structure.

Probabilistic models represent communities through a latent partition of the nodes that drives connection patterns. The stochastic block model (SBM) and its

extensions (Lee and Wilkinson, 2019) represent the canonical framework, assuming that the probability of an edge between two nodes depends solely on their community memberships. In networks with degree heterogeneity, the degree-corrected SBM (DC-SBM, Karrer and Newman, 2011) allows nodes within the same group to have different expected degrees, better reflecting real-world communities where hubs and peripheral nodes coexist. Hierarchical variants of the SBM explicitly model multi-layered organisation, enabling the detection of hierarchical community structures that arise from nested dense regions surrounded by sparser ones (Peixoto, 2014, 2019; Côme et al., 2021). Recent contributions integrate node attributes directly into the probabilistic framework (Stanley et al., 2019).

Spectral clustering (SC) and its extensions identify communities by partitioning the network through eigenvectors of the network's Laplacian matrix. Methods such as Covariate-Assisted SC (CASC) (Binkiewicz et al., 2017) and the Network-Adjusted Covariates (NAC) method (Hu and Wang, 2024) incorporate node attributes by modifying the similarity matrix or regularising structural information with attribute-based similarities. In particular, NAC aggregates covariate information from neighbours, weighting attributes by connection strength. This adjustment implicitly emphasises high-degree or influential nodes and has been shown to outperform the popular SC and CASC methods.

Density-based approaches interpret communities as dense regions separated by sparser areas, aligning directly with the intuitive definition of community structure. From a statistical perspective, this view is rooted in the modal formulation of clustering, where groups correspond to high-density regions of an underlying distribution, as discussed in Menardi (2016). Early adaptations such as DENGRAPH (Falkowski et al., 2007) extended DBSCAN (Ester et al., 1996) to graph data, identifying communities as locally dense neighbourhoods.

Leader-based methods represent a related but distinct paradigm: they assume that communities form around influential or central nodes, with other nodes gravitating toward these leaders based on structural proximity or attribute similarity. Early algorithms such as LICOD (Yakoubi and Kanawati, 2014), influence-propagation models (Helal et al., 2017), and eigenvector-based leader detection (Ahajjam et al., 2018) exemplify this idea.

A key advantage of DeCoDe approaches is that they accommodate heterogeneous cluster shapes and variable node density, thus the presence of multiple dense cores, making them suitable for networks exhibiting such structural patterns. An important recent contribution in this direction (Menardi and De Stefano, 2022) defines leaders as density peaks in the relations space and assigns remaining nodes to their nearest denser neighbour along topological paths.

Like most leader-based (or leader-first) methods, DeCoDe relies on structural measures to identify influential nodes. However, these measures alone may not fully capture social influence, particularly in networks where node attributes provide additional context. To address this, recent leader-based algorithms have incorporated attributive information alongside topology. For example, the aLBCD method (Helal

et al., 2016) assigns leaders based on attribute similarity when topological data is incomplete, effectively treating attributes as a complement to structural information. Similarly, TALB (Lu, 2021) incorporates attribute information before community detection by modifying the original network structure through a non-fixed weighted topology approach (Chunaev, 2020).

A key conceptual distinction in the leader identification step concerns the source of node popularity, which can emerge either endogenously–through preferential attachment processes within the network's structure–or exogenously, from actor attributes such as expertise, complementarity, or availability of resources (Barabási and Albert, 1999; Bianconi and Barabási, 2001; Papadopoulos et al., 2012). Most community detection methods, including DeCoDe, primarily capture endogenous popularity by identifying structurally central nodes that attract many links due to their position. However, in many social and collaborative systems, influence is also shaped by exogenous popularity: nodes become focal points not only because they are well-connected, but because they possess valuable or relevant characteristics that attract others. Accounting for this aspect enables a more realistic representation of community formation mechanisms, where connectivity emerges jointly from structural and attribute-driven affinity.

Probabilistic, spectral, and density-based approaches offer complementary views of community structure: the first rely on generative assumptions, the second on the global geometry of the graph, and the third emphasises local density and heterogeneous node influence. For networks such as Horizon collaborations–characterised by topic-specific subgroups and marked degree variability–the latter perspective is particularly interesting.

Chapter 3

Distance-based approach for density-based community detection

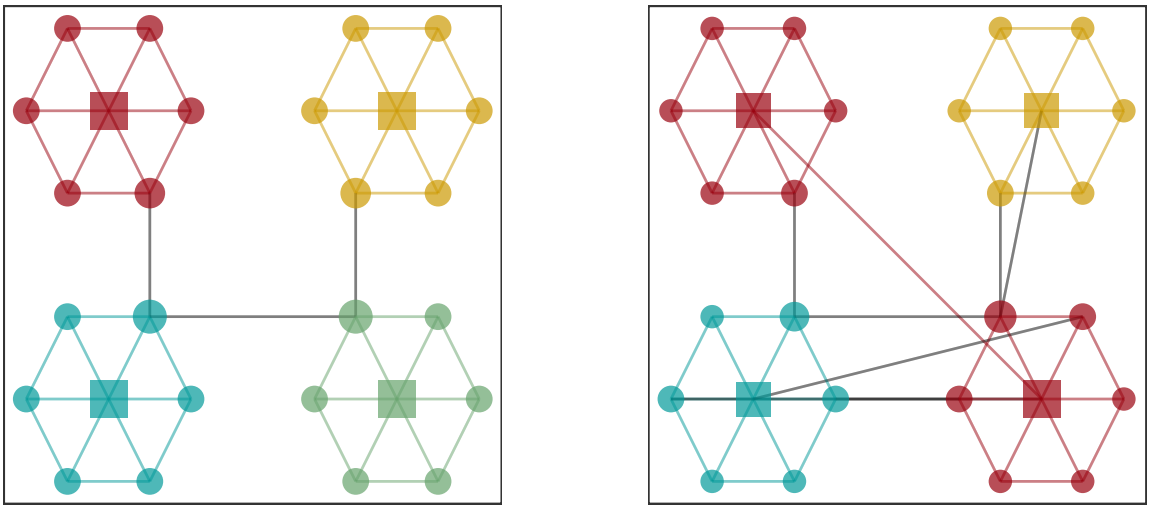
3.1 Motivating example

3.1.1 Structure-driven leader identification

In community-structured networks where leaders act as primary aggregation drivers, leadership is typically associated with nodes that exert strong local influence, i.e. those characterised by a high degree or located at the core of a tightly connected neighbourhood. This perspective aligns naturally with DeCoDe: when communities arise as dense subgraphs with limited inter-community interaction, the nodes with the highest number of ties within each group constitute its structural leaders. These leaders appear at the peaks of the node-wise density landscape and serve as anchors for the formation of cluster cores.

To illustrate this mechanism, consider the baseline toy network structure shown in Figure 3.1–taken from the original DeCoDe contribution–in which communities are internally cohesive and externally isolated. Node colours indicate community membership recovered by the density-based method, node size represents degree centrality, and node shape identifies whether a node acts as a community leader or a member. At low mixing, i.e., low frequency of inter-community links (Figure 3.1a), each community is an almost self-contained region of high intra-group density. Leaders (highlighted with squares in the plot) are the nodes with the highest degree centrality, and they occupy the most central positions within their respective dense subgraphs. In this ideal configuration, structural density and community organisation coincide perfectly: DeCoDe, built on degree, reliably identifies leaders as nodes lying atop these modal regions, and clustering simply follows the connectivity paths that link them.

However, real networks rarely exhibit such clean separations. As soon as inter-community links are introduced, structural prominence begins to shift in ways that complicate cluster cores identification. The proportion of cross-community edges in real networks tends to grow with node degree, thus highly connected nodes acquire more external ties than peripheral ones. In the toy network, nodes



(A) Lower mixing (B) Higher mixing

FIGURE 3.1: *Leader-based toy networks with (a) lower, and (b) higher levels of mixing. Node colours indicate density-based clusters, while node size reflects its degree centrality, and node shape its role as community leader or member.*

that were leaders under perfect assortativity retain the highest degrees even when mixing increases, since additional inter-community edges preferentially attach to these already central actors. This has important consequences: leaders become bridges rather than representatives of their own communities, and structural density shifts away from the actual community cores. As external edges accumulate, some nodes may appear structurally dense, not because they sit at the centre of a cohesive community, but because they mediate between multiple communities. Their high degree no longer reflects within-community cohesion but global popularity.

This effect is visible in the manipulated toy network with higher mixing in Figure 3.1b: leaders drift into positions that connect or even merge nodes belonging to what were initially distinct clusters (red cluster). Under such conditions, community recovery becomes challenging for any method that relies on structural density. Although the degree still correctly identifies the most connected nodes, these nodes cease to be informative about the underlying community organisation. Instead, they act as structural bridges whose presence erodes the separability of the true clusters.

These considerations, together with the acknowledgement that leader influence in social networks often interacts with homophily in shaping aggregation patterns, motivate the methodological extension introduced in Section 3.2. While structure-driven leader identification performs well in assortative networks with limited mixing, it rapidly loses reliability as the number of cross-community links increases. Our novel approach, termed Attributed Networks Density-based Community Detection (ANDeCoDe), is designed for settings in which homophily influences community formation. By modifying connection paths through a principled combination of

structural and attributive information, ANDeCoDe restores the interpretability of leader roles by preventing edges connecting attribute-dissimilar nodes from dominating the clustering process.

A further extension replaces the structural information based solely on direct neighbour relations with a structural similarity matrix, allowing the model to capture shared-neighbourhood patterns and thereby reducing the influence of high-degree bridge nodes on community detection.

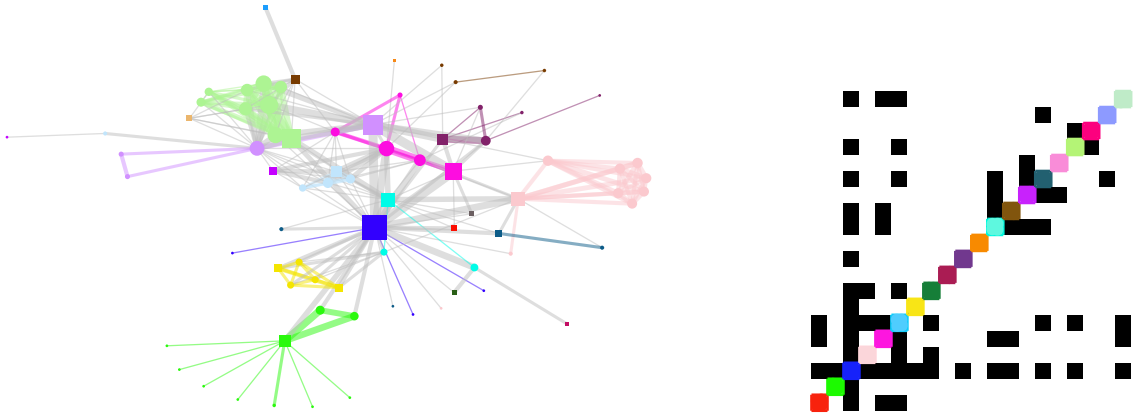
In summary, the structure-driven leader identification step in DeCoDe faces a fundamental limitation in networks with degree heterogeneity and mixing: structural prominence does not necessarily coincide with community representativeness. Recognising this distinction motivates the development of ANDeCoDe, which extends DeCoDe by integrating alternative structural similarity measures and attribute information in the community detection process, thereby improving the identifiability and interpretability of communities in realistic network configurations.

3.1.2 The Les Misérables network

To further illustrate the challenges arising in leader-based community detection and to motivate the methodological developments proposed in this work, we examine a well-known benchmark dataset in the network analysis literature: the Les Misérables character network (Figure 3.2). This network, originally introduced by Knuth (1993), encodes interactions among 77 characters in Victor Hugo's novel, where an edge between two characters indicates their co-appearance within one or more chapters, with edge weights indicating the frequency of co-appearance. The structure is archetypal of social networks that feature the interplay of complex community formation mechanisms, such as leader influence and homophily.

A key feature of this dataset is the availability of a meaningful reference partition. The community membership corresponds to the chapter in which each character first appears, yielding a clustering into 20 small groups. Although derived from narrative progression rather than structural properties, this partition reflects relational proximity in the storyline and is commonly used as a benchmark for evaluating community detection performance (as in the DeCoDe original work).

Figure 3.2 illustrates the structural organisation of the Les Misérables network. Panel (a) displays the full network; nodes are coloured according to their membership to the 20 reference communities, with node size proportional to degree centrality and node shape representing leaders, defined as the nodes with the maximum degree in their respective communities. Panel (b) displays the leader–leader connectivity matrix, where black cells indicate the presence of an edge between a given pair of leaders. This matrix reveals the extent to which leaders belonging to different communities are structurally linked, illustrating the potential for leader drifting and suggesting the need to incorporate attribute-driven similarity into the detection procedure.



(A) (B)

FIGURE 3.2: Les Misérables network. (a) Network partition into 20 reference communities derived from characters' first appearance in the novel. Node colours indicate true community membership; node size reflects degree centrality; and node shape represents leaders, defined as the nodes with the maximum degree in their respective communities. (b) Leaders' connectivity matrix: black cells denote the presence of an edge between two leaders in the original network.

In addition to the structural information, node-level attributes are indeed available. For each character, the dataset of Kalugin (2015) provides textual descriptions summarising narrative roles and social affiliations. To incorporate this information into our analysis, we transform the textual corpus into a tf–idf (Term Frequency–Inverse Document Frequency) matrix (Robinson, 2017), producing a numerical embedding of character profiles. This representation maps descriptive content into a numerical attribute space, enabling similarity relationships–such as common roles, alliances, or social and family affiliations–to be captured.

The combination of structural and attributive information makes the Les Misérables network particularly suitable for our purposes. On the one hand, the attribute information is rich enough to reveal homophilic tendencies aligned with narrative context. On the other hand, the network exhibits marked heterogeneity in degree, ranging from individuals who participate in only one or two interactions to highly central figures with more than a hundred connections. Jean Valjean (blue square in Figure 3.2), for example, interacts with 17 different characters, and the weighted degree–counting repeated co-appearances–reaches 158. This asymmetric distribution complicates the identification of cluster cores: central hubs can dominate the density landscape even when they are not representative of any specific community. Indeed, when the original DeCoDe algorithm is applied to the unweighted network of Les Misérables, using degree as a measure of density, the structure collapses into a single cluster, demonstrating how heterogeneity in degree masks the separation of communities in this context (see Section 3.4).

In summary, the dataset offers an empirically grounded example in which to observe how purely structural leader-based methods may fail to recover meaningful community structure, and how incorporating attribute-driven similarity can stabilise cluster cores identification and refine community boundaries.

3.2 Methodology

3.2.1 Attribute-driven density-based community detection

The clustering problem in social networks, as explored in Menardi and De Stefano (2022), can be approached by identifying high-density groups, where density is defined as a node-wise measure. Nodes are clustered when their density exceeds a threshold, and a path connects them. In the original DeCoDe formulation, node-wise density is a structural measure, and clusters emerge by following connection paths linking dense (modal) actors. As the density threshold varies, these connections generate a cluster tree that captures how high-density regions merge into broader community structures.

From the perspective of DeCoDe, the phenomenon of "leaders drifting" described in the previous section has concrete implications. Since cluster formation depends on identifying high-density nodes and linking them via connection paths, even a small number of cross-community edges can cause leaders from different communities to become mutually reachable at high density levels. This results in the early merging of branches in the DeCoDe cluster tree, thereby obscuring the true community structure. In other words, the algorithm continues to rely on accurately identified leaders, but these leaders no longer correspond to the cores of their respective clusters.

As the proposed ANDeCoDe algorithm builds upon the DeCoDe framework of Menardi and De Stefano (2022), further methodological details of the baseline procedure can be found in Algorithm 1. ANDeCoDemodifies the paths through which density-based clusters are formed. While density is still computed from structural properties, connectivity between nodes is altered through an attribute-informed similarity layer that penalises links between nodes dissimilar in the attribute space. The resulting procedure preserves the interpretability of density peaks while ensuring that paths are coherent with both the network topology and the attribute structure.

Topological and attribute information are combined through a unified distance-based approach that allows both the presence/absence of links and their weights to be adjusted when attribute information is taken into account. This modification of the network structure corresponds to what Chunaev (2020) classifies as a non-fixed topology method, i.e., an approach in which attribute similarity can alter the original topology before community detection is performed.

Let $G = (V, E, \boldsymbol{X})$ denote a node-attributed network, where $V = \{v_1, \ldots, v_i, \ldots v_n\}$ is the set of nodes; $E = \{e_{ij}\}$ is the set of (undirected) edges, with e_{ij} denoting either

the presence of a link or its weight (for unweighted networks $e_{ij} \in \{0, 1\}$), $i \neq j$; the $n \times n$ adjacency matrix encoding the e_{ij} elements is an equivalent representation of the network; $\boldsymbol{X} \in \mathbb{R}^{n \times p}$ is the attribute matrix whose i-th row $\boldsymbol{x}_i$ collects the p attributes associated with node v_i.

For each node v_i, let w_{ij}^S denote the structural similarity between v_i and its neighbour v_j, derived either from the edge set E or from a structural similarity measure. Notably, the original DeCoDe framework did not incorporate alternative structural similarity measures and relied solely on direct connections between nodes, encoded in the network's weighted (or unweighted) adjacency matrix. We extend this formulation by introducing the option to compute structural similarity using measures such as the Jaccard index, which quantifies the proportion of shared neighbours between adjacent nodes, making it sensitive to local clustering. Nodes at the centre of a well-defined community tend to share many neighbours with other members of that community. Under moderate levels of mixing, even if leaders drift outward, they typically retain a higher proportion of shared neighbours with their original community than with nodes in other groups.

Attribute similarity w_{ij}^A between node attributes $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ is computed via an appropriate distance-based measure and stored in the attribute similarity matrix W_A. W_A can be viewed as a fully connected weighted adjacency matrix. To consider exclusively nodes that are highly similar and mitigate the effect of weak or noisy attribute relationships, the matrix W_A is thresholded by retaining only entries whose similarity values exceed the 0.75 empirical quantile. The choice of the 0.75 quantile is justified in Section 3.3.2, where we show that the resulting partitions are stable and interpretable across a range of quantile thresholds. This operation filters out low-similarity pairs and yields a sparser, more informative similarity-weighted network.

We then modify the network structure by combining the normalised structural and attribute-aware matrices through a weighting parameter α:

$$w_{ij}^{\alpha} = \alpha \cdot w_{ij}^S + (1 - \alpha) \cdot w_{ij}^A,$$

where $\alpha \in [0, 1]$ governs the relative contribution of topological and attributive information. This produces a weighted attribute-aware network $G^{\alpha} = (V, E^{\alpha})$, with $e_{ij}^{\alpha} = w_{ij}^{\alpha}$, in which connectivity patterns can be reinforced or introduced when they align with attribute similarity and weakened or removed when they diverge from attribute similarity.

DeCoDe is then applied to the weighted network G_{α} using the *OR-based rule* for path construction introduced in Menardi and De Stefano (2022). Density is computed from a node-wise centrality measure, identifying the modal actors around which communities form. Each node v_i is then assigned a structural density estimate $\hat{\delta}(v_i)$. For each density threshold λ, the upper-level set $V_{\lambda} = \{v_i \in V : \hat{\delta}(v_i) \geq \lambda\}$ induces a subgraph $G_{\lambda} = (V_{\lambda}, E_{\lambda})$, constructed by retaining only the edges in G_{α} that provide the strongest incidence relation for each node. This ensures that connection paths remain coherent with both the structural and attribute information encoded in the

network. As λ decreases, the connected components of G_λ merge, generating a cluster tree; community cores correspond to the connected components that appear at the lowest levels of the tree. In addition to the detected number of clusters $\hat{K}$ and the cluster membership $C(v_i) \in \{1, \ldots, K\}$, the algorithm outputs an indicator $r(v_i)$ specifying whether node v_i acts as a cluster core ($r(v_i) = 1$) or as a non-core member ($r(v_i) = 0$). The resulting algorithm is summarised in Algorithm 1.

ANDeCoDe preserves the fundamental principle of density-based clustering, which consists of identifying modal nodes and connecting them through feasible paths, but modifies these paths so that they incorporate attribute similarity in addition to structural proximity. While structure-driven density continues to determine the leader nodes in each cluster, attribute-informed topology regulates how other nodes connect to these leaders. This prevents structurally induced but attribute-inconsistent connections from dominating the clustering process.

When $\alpha = 0$, topology information is used solely for computing node density (i.e., leader identification), while connectivity paths depend entirely on attribute similarity. When $\alpha \in]0, 1[$, attributive and structural information are jointly incorporated in defining the path structure. When $\alpha = 1$, only topology information is used, recovering the original DeCoDe formulation.

3.3 Simulation study

3.3.1 Simulation design

To evaluate the performance of the proposed density-based algorithm, we conduct two complementary simulation studies designed to compare the original DeCoDe method ($\alpha = 1$) with its attribute-driven extension, ANDeCoDe ($\alpha \neq 1$). The first focuses on a leader-based toy network and involves fine-grained control over structural and attribute-based parameters. The second employs a generative process to assess performance under varying network structural characteristics. The full details of this second simulation study are presented in Appendix D; the ideas guiding the construction of the desired structural patterns originate from the methodological framework developed in Chapter 5. Section 4.3.2 reports the performance of competing methods for attributed-networks community detection on simulated data.

Together, the two studies provide a comprehensive examination of how attribute similarity and network topology influence community detection performance. Our simulation framework makes a targeted modelling choice: attribute data are generated to perfectly align with the true community structure, while network topology is allowed to be disassortative through the introduction of inter-community mixing. This design reflects realistic scenarios where node attributes exhibit clear homophilous patterns, while the observed network structure is noisy due to structural constraints, measurement error, or non-homophilous attachment mechanisms (e.g., degree-based preferential attachment or institutional requirements).

The agreement between the inferred and the true cluster partitions is quantified using the Normalised Mutual Information (NMI, Danon et al., 2005), which ranges from 0 (no agreement) to 1 (perfect agreement). NMI is an information-theoretic measure that quantifies how much information is shared between two partitions,

Algorithm 1 ANDeCoDe

Input:
Attributed network $G = (V, E, X)$
Weight parameter $\alpha \in [0, 1]$

Step 1: Structure-driven leader identification
Compute structural node-wise densities $\hat{\delta}(v_1), \ldots, \hat{\delta}(v_i), \ldots, \hat{\delta}(v_n)$

Step 2: Construction of the attribute-aware network
Compute the structural similarity matrix W_S from E
Compute the attribute similarity matrix W_A from X
Sparsify W_A by retaining entries above the 0.75 quantile
Combine structural and attribute similarities:
$w_{ij}^{\alpha} = \alpha \cdot w_{ij}^{S} + (1 - \alpha) \cdot w_{ij}^{A}$
Let $G^{\alpha} = (V, E^{\alpha})$ be the resulting weighted attribute-aware network, with $e_{ij}^{\alpha} = w_{ij}^{\alpha}$

Step 3: Density-based community detection
For each node v_i, identify its maximum-weight incident edge:
$\tilde{e}_{im}^{\alpha} = \max_{j: e_{ij}^{\alpha} \in E_{\alpha}} e_{ij}^{\alpha}$
for $0 < \lambda < \max_i \hat{\delta}(v_i)$ **do**
  Determine the upper-level set:
  $V_{\lambda} = \{v_i \in V : \hat{\delta}(v_i) \geq \lambda\}$
  Build the induced graph $G_{\lambda} = (V_{\lambda}, E_{\lambda})$, where
  $$E_{\lambda} = \left\{ e_{ij}^{\alpha}(\lambda) : e_{ij}^{\alpha}(\lambda) = \begin{cases} e_{ij}^{\alpha} & \text{if } v_i, v_j \in V_{\lambda} \text{ and } (\tilde{e}_{im}^{\alpha} = e_{ij}^{\alpha} \text{ or } \tilde{e}_{jm}^{\alpha} = e_{ij}^{\alpha}), \\ 0 & \text{otherwise} \end{cases} \right\}$$
  Identify the connected components of G_{λ}
  Update $\tilde{e}_{im}^{\alpha} = \max_{j: e_{ij}^{\alpha} \in E^{\alpha} \setminus E_{\lambda}} e_{ij}^{\alpha}$
end for
Build the cluster tree associating each λ to the number of connected components of G_{λ}
Identify cluster cores as the connected components of G_{λ} at the lowest λ levels for which the branches of the tree represent the leaves

Output:
Role $r(v_i) \in \{\text{core}, \text{member}\}$
Cluster membership $C(v_i) \in \{1, \ldots, K\}$
Number of detected clusters $\hat{K}$

with values closer to one indicating stronger correspondence. It is widely used in community-detection studies because it provides an intuitive measure of overlap between inferred and reference cluster assignments. We adopt NMI here both because it provides an interpretable measure of partition similarity and because it was the primary evaluation metric in the original DeCoDe work, facilitating direct comparison with our extension.

Simulation Study 1: Leader-based toy network The first simulation study aims to investigate how attribute similarity and mixing between communities affect the performance of DeCoDe and ANDeCoDe. We use a leader-based toy network (Figure 3.1) whose community structure is shaped by two social mechanisms: leader attraction and homophily. This network provides a representative setting to compare the behaviour of the algorithm across different attribute types and across a fine grid of tuning parameters.

The true community membership is defined by a fixed leader-driven configuration. Mixing across communities is controlled by the global parameter $\mu \in [0, 1]$, which regulates the proportion of links that connect nodes outside their own community (Lancichinetti et al., 2008). Increasing μ increases community overlap, as each node shares a fraction $1 - \mu$ of its links within its community and μ with nodes in external communities. When $\mu > 0.5$, each node maintains more connections outside its community than within it, and the community structure becomes effectively undetectable.

To examine the impact of attribute nature on performance, we generate different types of homophilous node attributes, all sharing the same underlying cluster structure as the communities:

- **Continuous attributes**: bivariate Gaussian and Student-t distributions;
- **Binary attributes**: two independent Bernoulli variables;
- **Mixed attributes**: combinations of Gaussian and binary variables.

Attribute similarity is computed using the Euclidean distance for continuous data, the simple matching coefficient for binary data, and the Gower distance for mixed data Gower (1971).

For each attribute type, networks are generated for μ varying from 0 to 1 in increments of 0.1. For each value of μ, we simulate 100 independent replicates. ANDeCoDe is then applied using α varying from 0 to 1 in increments of 0.1, allowing the evaluation of its performance across a dense grid of attribute–structure weightings.

This experiment provides insight into: how strongly attribute similarity drives clustering accuracy; how the balance between structural and attribute similarities α interacts with increasing inter-community mixing μ; how the nature of the attribute space (continuous, binary, mixed) affects performance.

Simulation Study 2: DC-SBM Networks The second simulation study investigates the ability of ANDeCoDe to recover density-based communities under structural variability. This experiment focuses on how performance changes with network size, number of communities K, and community size imbalance.

Networks of size $n \in \{50, 150, 300\}$ and partitioned into $K \in \{5, 10\}$ clusters are generated using DC-SBM (Karrer and Newman, 2011), a variant of the SBM that incorporates node-specific degree parameters to reproduce heterogeneity in connectivity and account for preferential attachment. Community assignments are obtained from either a uniform distribution (equal-sized communities) or a Dirichlet distribution (introducing size imbalance). Mixing between communities is controlled by the parameter μ. Node attributes are sampled from 2D Gaussian mixture models (Bouveyron et al., 2019) with spherical covariance matrices, generated so that their cluster structure matches the network community structure.

For each scenario, combined with each mixing level $\mu \in \{0, 0.2, 0.4\}$, we generate 100 independent network realisations. Here, $\mu = 0$ signifies the absence of inter-community links, whereas in the previous example, it indicated that no additional inter-community links were introduced. Further details on the data generation mechanism are provided in Section 4.3.1. ANDeCoDe is then applied to the resulting networks using $\alpha \in \{0, 0.25, 0.5, 0.75, 1\}$.

This experiment provides insight into: how community recovery is influenced by structural scale, including network size and number of communities; how strongly mixing and cluster size imbalance undermine community detection performance; how the balance between structural and attribute similarities α interacts with increasing inter-community mixing μ under realistic network variability.

3.3.2 Simulation results

Simulation Study 1: Leader-based toy network Figure 3.3 displays the NMI median scores evaluated on the manipulated toy networks across parameter values and attribute types. Results indicate that ANDeCoDe's performance generally decreases as the mixing level μ increases, reflecting the challenge of detecting communities with more inter-community connections. Notably, the DeCoDe ($\alpha = 1$) original algorithm's NMI values drop significantly with just 0.1 mixing. Performance varies with the α parameter, where combining attribute and structural information achieves the best results in the presence of mixing. The Jaccard similarity matrix shows greater stability, particularly for binary data, suggesting that it is more stable in noisy network configurations. Overall, in contrast to DeCoDe, ANDeCoDe for $\alpha < 1$ leads to enhanced community detection performance in networks with homophily on observed attributes.

We also assessed the stability of the algorithm's performance, quantified using the Median Absolute Deviation (MAD) of the NMI scores across network replicates. Overall, the MAD values were low (median = 0.05; mean = 0.07), with many parameter combinations exhibiting a MAD of 0, indicating highly consistent performance across

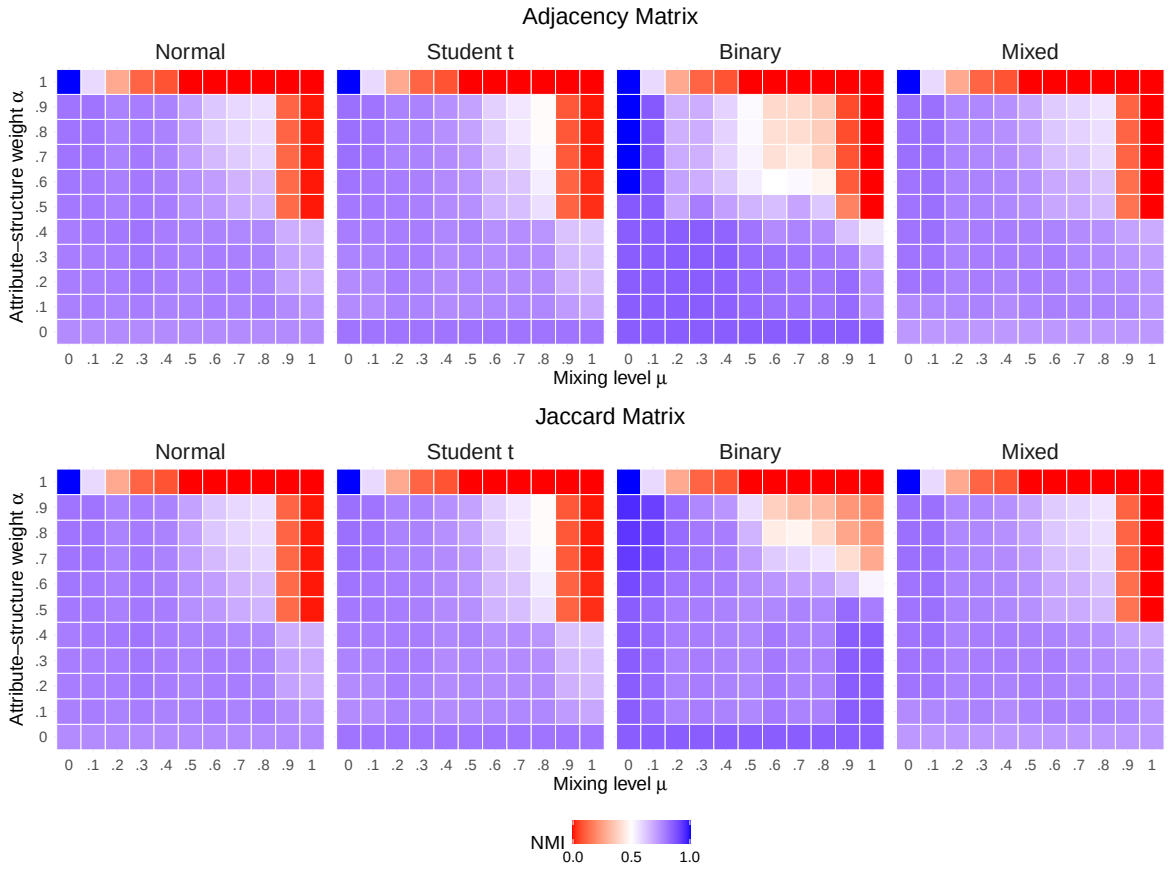


FIGURE 3.3: Normalised Mutual Information (NMI) for different values of weight α and mixing parameter μ. Performance on the toy network is compared between adjacency-based and Jaccard-based approaches and tested across different attribute types, including Gaussian, Student's t-distributed, binary, and mixed data. High median NMI values (computed over network replicates) are in blue, low values in red.

replicates. Notable variability was observed only in a few specific configurations (e.g., $\alpha > .5$ and binary data), where MAD values reached approximately 0.3.

To complement the simulation study, we conducted a sensitivity analysis to evaluate the robustness of using the 0.75 quantile as the threshold for sparsifying the attribute similarity matrix W_A. Specifically, we assessed how variations in the threshold impact the performance of the DeCoDe algorithm. We considered a range of quantile thresholds: from 0.5 to 0.75, with increments of 0.05. Thresholds above 0.75 were excluded, as they resulted in excessive sparsification, thus loss of meaningful relationship information. For each threshold, we evaluated clustering performance across different values of μ, α, and attribute types. The results show that ANDeCoDe's NMI values are generally stable across the considered quantile range. In particular, the MAD remained lower than 0.1 for 97.3% of parameter combinations, indicating

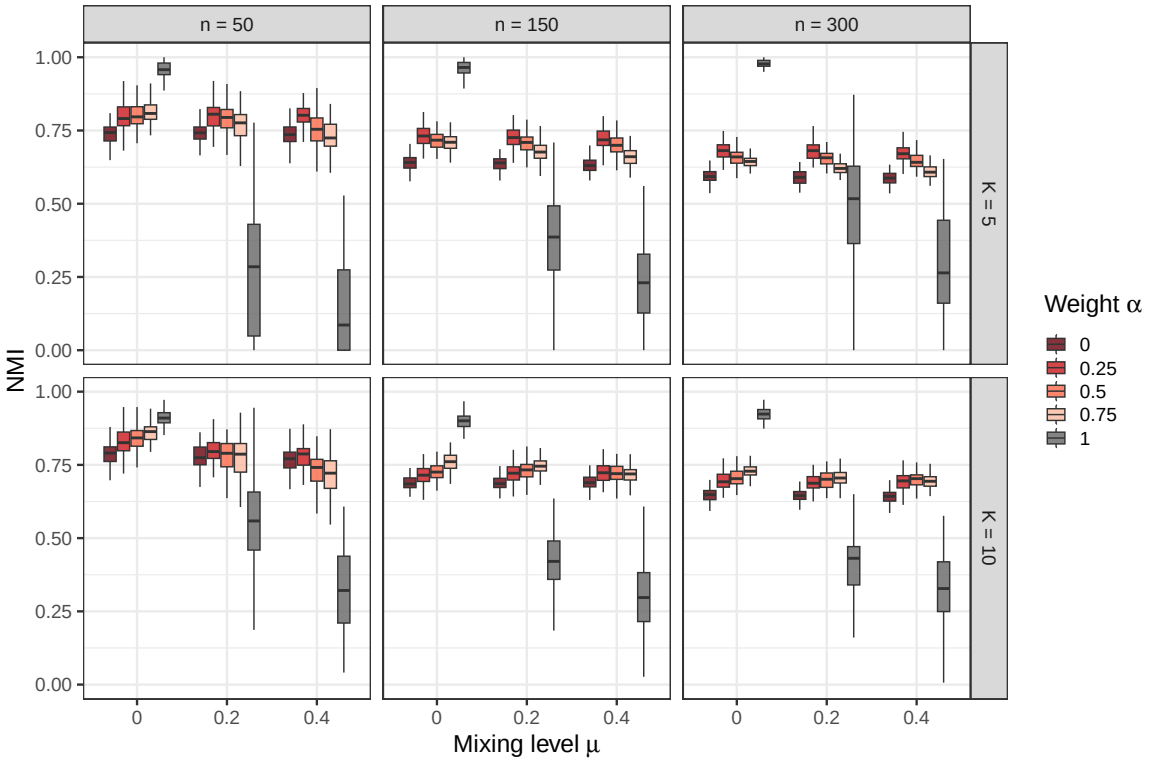


FIGURE 3.4: ***Adjacency matrix*** *– Normalised Mutual Information (NMI) distribution across different numbers of nodes n, numbers of communities K, mixing levels μ, and attribute–structure weights α. Results are shown for simulated networks with uniform community sizes.*

that small changes in the sparsification threshold do not influence outcomes in these configurations. For values of α around 0.6 and high values of mixing μ, we observed a slight increase in MAD (ranging from 0.21 to 0.24), suggesting some sensitivity but still within acceptable limits. These findings suggest that the 0.75 quantile offers a practical trade-off between retaining meaningful attribute similarities and avoiding excessive noise from weak connections.

Simulation Study 2: DC-SBM Networks Figure 3.4 shows the distribution of NMI scores obtained for ANDeCoDe with the adjacency-based structural similarity on simulated networks, across all combinations of network size n, number of communities K, mixing level μ, and attribute–structure weight α. Several patterns emerge from these results.

NMI values remain relatively stable across different values of n and K, with smaller networks exhibiting slightly more variability. For DeCoDe ($\alpha = 1$), performance consistently decreases as the mixing parameter μ increases, reflecting the expected loss of community separability in DC–SBM networks with denser cross-community connectivity. Notably, this effect arises almost exclusively when clustering relies solely

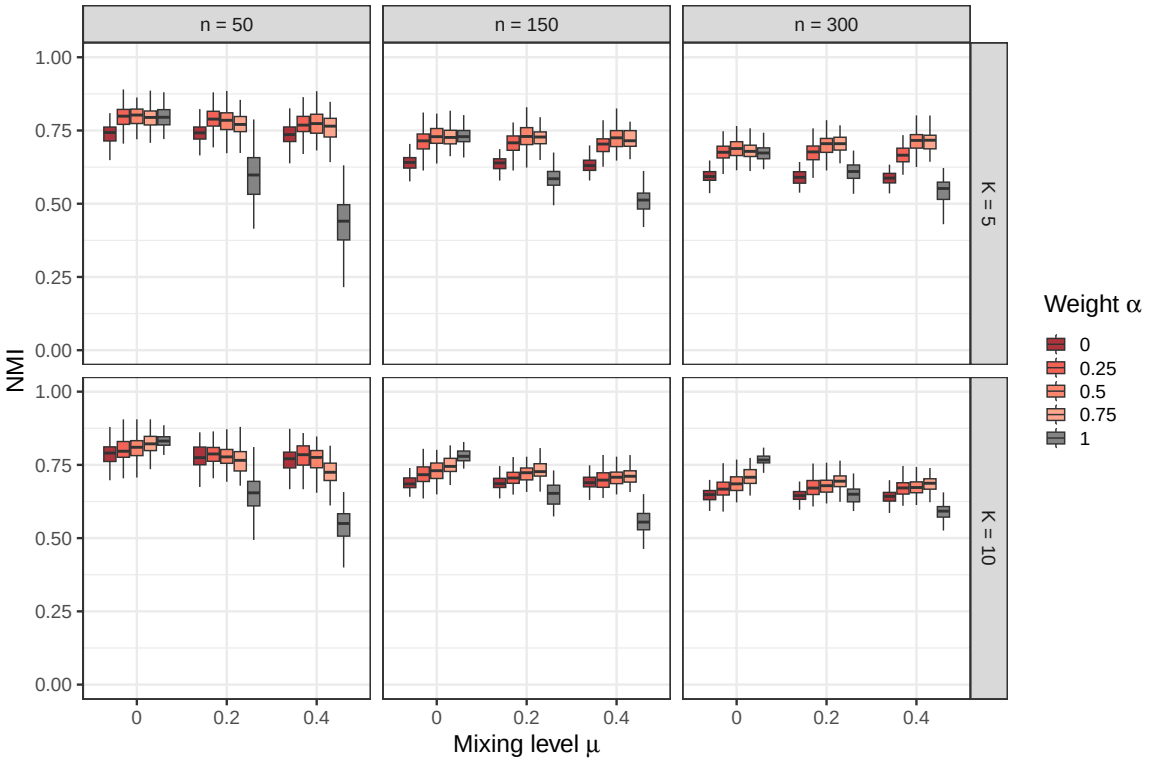


FIGURE 3.5: ***Jaccard matrix*** *– Normalised Mutual Information (NMI) distribution across different numbers of nodes n, numbers of communities K, mixing levels μ, and attribute–structure weights α. Results are shown for simulated networks with uniform community sizes.*

on structural information, as in the original DeCoDe. As shown by the grey boxplots in Figure 3.4, dependence on the raw adjacency structure results in substantial performance degradation even at moderate levels of mixing $\mu = 0.2$. This mirrors the behaviour observed in Section 3.1: structural information alone is insufficient when mixing produces leader drifting. Conversely, in networks with no inter-community connections ($\mu = 0$), relying on attribute similarity may introduce weak mixing that was not present in the topology, thereby slightly reducing performance. As a result, the original DeCoDe performs best under ideal, but unrealistic, conditions in which communities are perfectly assortative and no mixing occurs.

Compared to DeCoDe ($\alpha = 1$), ANDeCoDe with intermediate values of α (0.25–0.5) consistently produces the highest NMI across all network configurations in the presence of mixing. These settings balance structural and attribute information. In contrast, when $\alpha \approx 0$, performance becomes more sensitive to mixing, indicating that moderate attributive weighting is beneficial, but too much emphasis on the filtered attribute similarity matrix W_A leads to a loss of important structural information.

These findings align with those from Simulation Study 1, where the attribute-structure integration, as in ANDeCoDe, was essential to enhance density-based

community detection performance in networks with moderate to high mixing.

The results for the topology-only DeCoDe ($\alpha = 1$) with Jaccard-based approach in Figure 3.5 show greater robustness to mixing. When $\mu = 0.2$ or $\mu = 0.4$, the Jaccard-based method retains substantially higher NMI values than the adjacency-based approach. This confirms that common-neighbour similarity provides a more reliable structural information in networks with degree heterogeneity, as anticipated in Section 3.2. More clearly than in the adjacency case, performance under Jaccard similarity increases with network size, and the performance gaps between mixing levels become smaller in larger networks.

Across all conditions, the highest NMI values under mixing are typically achieved for ANDeCoDe with $\alpha \in [0.5, 0.75]$. This means that scenarios giving greater weight to the structural component tend to perform better, indicating that the Jaccard-based structural similarity provides a more informative representation of the underlying topology than direct-neighbourhood information alone, and thus contributes more effectively to identifying community structure when mixing is present.

Taken together, Simulation Studies 1 and 2 show that ANDeCoDe, by incorporating attribute information, enables meaningful community structure identification when degree heterogeneity and mixing are present. The results further highlight that substituting the adjacency matrix with a Jaccard similarity matrix enhances model strength by aligning the structural information with local cohesion rather than degree distribution. Across all network configurations, a moderate level of attribute–structure integration consistently yields the best trade-off between the two sources of information.

3.4 Application: the Les Misérables network

A case study involves examining how our method performs on the popular Les Misérables network, where community membership and attributes of nodes are available. We used the textual descriptions of characters as node attributes. In Figure 3.6, attributive edges are highlighted in red and topology edges, representing characters' co-appearance, are shown in grey. Each block in the adjacency matrix corresponds to a chapter, and block elements indicate characters who early co-appeared in that chapter.

Table 3.1 reports clustering performance on both the unweighted and weighted versions of the benchmark network-—where edge weights reflect the frequency of character co-appearances—-evaluated across a range of values for α and under alternative specifications of structural similarity. Agreement with the early-appearance classification is assessed using NMI, Adjusted Rand Index (ARI), and the number of recovered clusters K. The ARI complements NMI by evaluating pairwise agreement between partitions while explicitly correcting for agreement that could arise by chance. ARI values equal to 1 indicate perfect agreement, values near 0 correspond to chance-level similarity, and negative values indicate less agreement than would be expected at random (Hubert and Arabie, 1985a). This adjustment makes ARI

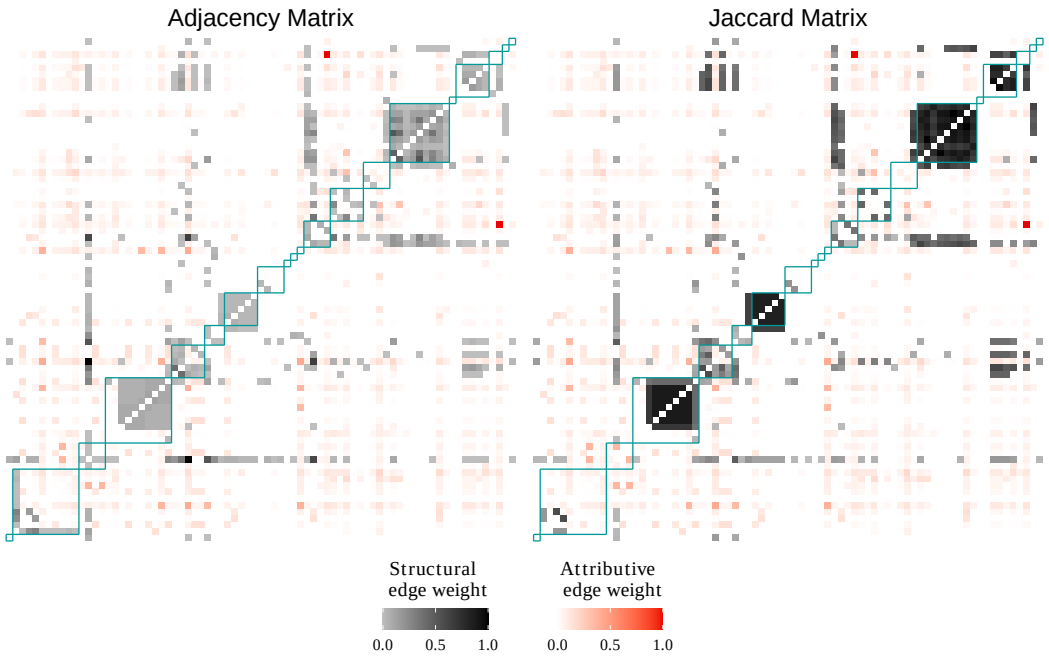


FIGURE 3.6: Les Misérables character network. Attributive edges are shown in red and structural edges in grey, prior to their weighted combination. Each block, outlined by blue lines, corresponds to a chapter, and block elements indicate characters who early co-appeared in that chapter.

particularly useful when comparing clustering outcomes across different numbers or sizes of communities, where unadjusted measures may be misleading.

The attributive-only setting ($\alpha = 0$) yields moderate NMI (≈ 0.45–0.47) but essentially no agreement in ARI, while substantially overestimating the number of clusters ($\hat{K} \approx 25$–27). This indicates that attributive information alone is insufficient to reconstruct the structural organisation of the network and tends to produce fragmented partitions.

ANDeCoDe, by introducing structural information ($\alpha > 0$), improves performance across all measures. For the adjacency-based approaches, balanced or topology-dominant weights ($\alpha = 0.50$ or 0.75) substantially increase both NMI and ARI, with the weighted adjacency achieving NMI values up to 0.66 and ARI up to 0.35. However, adjacency-based methods remain sensitive to degree heterogeneity: in the unweighted case, even the best-performing configurations recover only $\hat{K} = 5$ clusters, well below the reference $K = 20$, reflecting a tendency to merge structurally cohesive but topologically proximate groups.

The Jaccard-based approaches show a different–and improved–performance profile. Under topology-dominant weighting ($\alpha = 0.75$), the unweighted Jaccard matrix achieves the highest NMI overall (0.67) and an ARI of 0.35, while recovering a number of clusters ($\hat{K} = 18$) close to the reference value. The weighted Jaccard matrix performs similarly well, with NMI around 0.59 and estimated number of clusters

between 19 and 21, further confirming its strength.

Taken together, these results highlight three key insights. First, attributive information alone leads to over-fragmentation, while structure alone, as in DeCoDe ($\alpha = 1$), risks excessive merging–particularly for unweighted adjacency, where the method degenerates to a single cluster. Second, ANDeCoDe, by combining attributes with structural similarity, markedly improves performance over the original DeCoDe approach, especially when α places moderate to strong weight on topology. Third, Jaccard similarity provides a more reliable structural representation than adjacency, yielding higher agreement with the reference partition and more accurate estimates of the number of communities.

3.5 Final remarks

The present study enhances the DeCoDe leader-based method by introducing ANDeCoDe, a novel framework which integrates structural and attributive information in density-based community detection. Our approach aims to identify communities in social networks where homophily and leader influence are key mechanisms. An evaluation using both simulated and real-world data suggests that integrating attribute information improves the determination of community boundaries, particularly in

TABLE 3.1: Performance of ANDeCoDe on Les Misérables benchmark network under different structural similarity measures (adjacency vs. Jaccard), weighting schemes (unweighted vs. weighted networks), and attribute–structure weights α. Bold values indicate the best performance within each block.

	Unweighted Adjacency			Weighted Adjacency		
α	NMI	ARI	$\hat{K}$	NMI	ARI	$\hat{K}$
0	0.47	0.00	**27**	0.45	-0.01	**25**
0.25	0.31	0.06	5	0.44	0.11	7
0.5	0.46	0.16	5	0.59	0.24	7
0.75	**0.47**	**0.18**	5	**0.66**	**0.35**	7
1	0.00	0.00	1	0.65	0.34	6
	Unweighted Jaccard			Weighted Jaccard		
α	NMI	ARI	$\hat{K}$	NMI	ARI	$\hat{K}$
0	0.47	0.00	27	0.45	-0.01	25
0.25	0.35	-0.01	15	0.40	0.01	17
0.5	0.43	0.08	14	0.51	0.06	17
0.75	**0.67**	**0.35**	**18**	**0.59**	0.13	**19**
1	0.60	0.23	13	**0.59**	**0.15**	**21**

networks with moderate to high levels of mixing. The optimal weight parameter α varies depending on the network structure and attribute data type, with values $0 < \alpha < 1$ usually yielding superior results in the presence of mixing. This suggests the choice for a balanced approach, with α close to 0.5. The parameter was manually tuned in the experiments, and no automatic approach was proposed for its selection in real applications. The development of a data-driven strategy for tuning α remains an important direction for future research.

Moreover, the choice of structural similarity measure plays a decisive role in performance. The Jaccard-based approach consistently exhibits greater stability, particularly in binary and noisy network settings, where shared-neighbourhood information provides a more reliable topological signal than direct adjacency. At the same time, the results highlight the limitations of relying exclusively on either structural or attribute information, especially in networks characterised by overlapping or weakly defined communities.

The variability observed across weighting schemes and similarity measures underscores the importance of carefully assessing how structural and attributive information should be combined for effective community detection. Both the choice of node-level structural similarity and the balance between topological and attribute-based contributions have a substantial impact on clustering performance. This points to the need for a tailored approach that adapts to the properties of the data, the underlying network characteristics, and theoretical assumptions about the mechanisms driving community formation.

By preserving essential topological features while integrating meaningful attribute information, ANDeCoDe offers a flexible framework for community detection in attributed networks. The next chapter further develops this line of work by introducing an alternative attribute-driven extension of DeCoDe. This contribution shifts the identification of leaders from purely structural centrality to high density in attribute space, allowing community structure to emerge around nodes that are representative in terms of their characteristics.

Chapter 4

Attribute-driven density estimation for density-based community detection

4.1 Motivating example

During the last decades, complex networks have been recognised to be crucial in explaining social phenomena using the structural and relational features of the network of actors involved. In the innovation economics and related public interventions, knowledge and innovation networks are currently of great interest.

Social network analysis has been recognised as a fundamental tool in studying inter-organisational interaction in terms of knowledge and innovation flows (Ter Wal and Boschma, 2009). In this stream of literature, network analysis tools have been adopted to explore EU-funded collaborative Research & Development (R&D) networks, derived from Horizon 2020 and Horizon Europe as key European initiatives supporting research and innovation. Data on research topics, objectives, organisations and other details on Horizon projects are available through the open access portal Community Research and Development Information Service (CORDIS) (European Commission, 2024).

From the CORDIS dataset, scientific collaboration networks can be constructed, where vertices represent participating organisations, and links correspond to EU-funded research projects in which they are involved. By mapping these connections, researchers can analyse how knowledge is transferred, innovations emerge, and technological capabilities develop across different sectors and countries. Such networking is often referred to as the Triple Helix model (Etzkowitz and Leydesdorff, 1995).

The topology and the community structure of these networks reflect not just scientific interactions, but also broader EU innovation strategies, policy impacts, and technological diversification efforts. Studies examined network structures in relation to EU policy objectives (Breschi and Cusmano, 2004), dynamics of innovation (Balland et al., 2019a,b; Maruccia et al., 2020), and the importance of technological diversity (Muscio et al., 2022). Barber et al. (2006) were among the first to analyse a large European R&D network, focusing on its topological properties. However, a limited number of studies have applied community detection techniques to networks derived

from such data, explicitly aiming to capture the tendency of specific organisations or regions to cluster together due to one or more underlying aggregation mechanisms. Notable exceptions include Barber and Scherngell (2013), who analysed the bipartite network of organisations and projects funded by the Fifth European Framework Programme using a variation of the label propagation algorithm, identifying relevant substructures characterised by spatially heterogeneous community groups. More recently, Genova et al. (2024), starting from a multipartite network of projects and using a community detection algorithm, identified field-specific collaboration clusters where some European countries play a central role in project participation. Another study, (Kosztyán et al., 2024) introduces a novel link prediction model for analysing Horizon 2020 collaboration networks, while also identifying communities and assessing their regional distribution. Although not explicitly a community detection method but rather a clustering approach, Cerqueti et al. (2024) cluster yearly EU-funded research collaboration networks by combining nodal centrality measures with rank–size analyses, enabling comparisons across years based on the roles of institutions and highlighting both similarities and differences in leading hubs and overall network structure. The presence of attributes, including scientific disciplines and project topics, guided the identification of distinct thematic sub-networks. Morea et al. (2024) and Morea and De Stefano (2023) analysed specific thematic Horizon networks through centrality measures and community detection.

All these studies have shown, in the European context, that these networks are crucial for understanding how transnational research collaborations contribute to competitive innovation ecosystems. The analysis of such data offers comprehensive information on how academic and research institutions, private companies, and public organisations across Europe collaborate to achieve highly innovative outputs.

This work examines EU-funded hydrogen projects, not only for their strategic importance in the energy transition, but because they provide an ideal context for analysing the dynamics through which collaborative innovation networks are formed and structured (Morea and De Stefano, 2023). Hydrogen valleys are integrated regional ecosystems in which research, industry, public authorities, and civil society work together to develop hydrogen-based solutions. These projects are designed to stimulate interaction between heterogeneous actors and promote innovation through a systemic approach.

Our objective is to analyse the relational structures that develop among organisations involved in Horizon hydrogen projects from 2015 to 2029. Previous analysis on the same dataset demonstrated that the overall knowledge flow is shaped by some pivotal organisations (Morea et al., 2024). As such, hydrogen valleys provide an exemplary case for studying leader-influence mechanisms that drive the formation of research communities.

4.1.1 The hydrogen Horizon network

The hydrogen Horizon network, bipartite in structure, describes the interactions between 859 organisations collaborating on 181 EU-funded projects, yielding a sparse one-mode projection with a density of 0.02%. For this study, we focus on the largest connected component, which contains 825 organisations. The 7,807 edges in the organisation one-mode network represent relationships derived from the shared connections in the project-organisation bipartite network. An edge between two organisations is formed if both are involved in at least one common project; the edges can carry weights that reflect the number of shared projects between two organisations. In this setting, edge weights are highly heterogeneous but also strongly driven by project size and consortia composition rules, which can inflate ties of large, central organisations that participate in many consortia. As a result, high weights often reflect prolific participation or administrative roles rather than meaningful leadership or homophilous attraction. Using such weights directly in a density-based clustering framework may bias the identification of clusters by obscuring attribute-driven cohesion. For this reason, in this work, we do not interpret edge weights as indicators of stronger connection paths in the clustering process.

Available information includes additional organisation-level and project-level attributes, i.e., geographical location, activity type, and thematic focus. This allows us to investigate whether geographical or cultural proximity, shared research interests, and institutional roles (research institutes, universities, industry, etc.) shape collaboration patterns, leading to the formation of subgroups by characteristics. At the same time, it is worth mentioning that Horizon programme rules encourage diversity within consortia: funded collaborations are expected to include partners from different countries and to promote interaction across activity types, particularly between academia and industry.

Table 4.1 reports the distribution of organisations by country and activity type. From the country perspective, the largest shares come from France (13.2%), Italy (12.1%), and Germany (10.9%), followed by Spain and the Netherlands. These countries represent major hubs within the European research area and coordinate a substantial proportion of Horizon-funded projects. Northern and Eastern European countries appear with smaller proportions, while the category Rest of Europe (RoE) aggregates several countries with individually small contributions. Overall, the distribution reflects the concentration of research capacity and Horizon participation in Western and Southern Europe.

With respect to activity type, private for-profit organisations (PRC) dominate the network (62.4%). Higher education institutions (HES) and research organisations (RECs) each account for approximately 13% of participants, forming the academic and scientific core of the network, while public organisations (PUB) and other entities (OTH) contribute less. This composition underscores the heterogeneous, multi-actor nature of Horizon consortia, which integrate universities, industry, and research centres in mission-oriented collaborations.

To better understand how research interests shape collaborations, we complement organisational characteristics with information on the scientific content of the projects. Each funded project is described by a list of keywords encoded using the European Science Vocabulary (EuroSciVoc, Publications Office of the European Union, 2025).

TABLE 4.1: Distribution of organisations involved in hydrogen Horizon projects by country, activity type, and dominant research topics.

	Level	N	%
Country	FR	109	13.2
	Italy (IT)	100	12.1
	Germany (DE)	90	10.9
	Spain (ES)	71	8.6
	Netherlands (NL)	69	8.4
	Norway (NO)	41	5.0
	Belgium (BE)	37	4.5
	Greece (EL)	37	4.5
	United Kingdom (UK)	34	4.1
	Denmark (DK)	28	3.4
	Finland (FI)	23	2.8
	Sweeden (SE)	17	2.1
	Austria (AT)	15	1.8
	Rest of Europe (RoE)	< 15	18.7
Activity type	PRC	515	62.4
	HES	113	13.7
	REC	105	12.7
	OTH	61	7.4
	PUB	28	3.4
	Missing	3	0.4
Top 10 keywords	fuel cells	316	38.3
	electrolysis	225	27.3
	natural gas	111	13.5
	ecosystem	110	13.3
	wind power	101	12.2
	catalysis	88	10.7
	sustainable economy	87	10.5
	business models	86	10.4
	energy conversion	67	8.1
	coating and films	58	7

We summarise keyword frequencies both at the project level (how often a keyword appears across projects) and at the organisation level (how many organisations are associated with each keyword through their project participation), thus distinguishing research topic prevalence from organisational thematic reach.

At the project level, the most frequent keywords reflect the thematic focus of the EU-funded hydrogen research: *fuel cells* and *electrolysis* (54 projects each), *catalysis* (20 projects), and *natural gas* (17 projects). At the organisation level, keyword frequencies reflect thematic reach. *Fuel cells* and *electrolysis* dominate, involving 38.3% and 27.3% of organisations respectively, confirming their central role in the network. *Natural gas*, *ecosystem*, *wind energy* (around 12-14%) indicate widespread engagement in adjacent energy sectors. The presence of *catalysis*, *sustainable economy*, and *business models* (around 10%) highlights the spread of socio-economic themes across many organisations, while more specialised areas are limited to smaller subgroups of experts.

Fifty projects contain one or more keywords that do not appear elsewhere, and only four projects have more than half of their keywords classified as unique. Examples include *simulation software*, *mathematical models*, and *thermodynamics* in project 101053133; *metallurgy* and *thermodynamic engineering* in 101091456; *tissue engineering*, *stem cells*, and *piezoelectrics* in 641640; and *crystallography*, *bioinorganic chemistry*, *quantum computers*, and *enzymes* in 745702. These cases highlight the presence of highly specialised research niches embedded within an otherwise thematically coherent research.

The initial observations regarding edge weights are consistent with the concentration of both attributes and connectivity in the network, which displays a markedly heterogeneous degree distribution characterised by a small number of highly connected hubs and a large majority of peripheral organisations. Degree ranges from 1 to 170, with a median of 15 and an upper quartile of 24, while only a very small fraction of nodes (fewer than 5%) attain a degree above 50. These hubs are disproportionately composed of private for-profit companies (PRCs), organisations based in Finland, and actors whose dominant thematic focus is *natural gas*, indicating that structural prominence may be closely associated with institutional roles, national participation patterns, and research domains characterised by lower entry barriers and widespread accessibility within the application context.

Collaborative ties in research networks appear to be shaped by the interplay of thematic priorities and leading organisations. Although this mechanism is not directly observable from the network structure alone, it seems to appear as a key aspect of bipartite event-actor networks, where the characteristics of the events that generate ties play a key role in identifying potential leaders.

In our application, project topics naturally encode node attributes: an organisation's thematic profile is derived from the collection of textual descriptors associated with the projects in which it participates. This representation facilitates the identification of similarity-driven attachment through homophily in the attribute space, consistent with popularity–similarity models of network growth (Papadopoulos et al., 2012). At

the same time, the same project-level information also captures structural popularity: organisations involved in many projects accumulate broader and more heterogeneous keyword profiles, making them more likely to occupy central regions of the attribute space and to be representative of a wide range of other actors. In this bipartite setting, similarity and popularity are therefore not independent dimensions but are jointly shaped by the underlying structure.

4.2 Methodology

4.2.1 Attribute-driven leader identification

A central challenge in social network analysis is the identification of influential individuals–often referred to as leaders–who shape the structure and evolution of communities. Traditional approaches to leader detection rely primarily on structural metrics such as degree, betweenness, or eigenvector centrality (Newman, 2010). While these measures capture positional importance in the network topology, they overlook a crucial dimension of social interaction: the attributes of individuals. In many real-world networks, homophily plays a fundamental role in shaping community boundaries.

As we noted in Chapter 3, this limitation is particularly evident in large collaborative networks, where structural importance can be misleading. Highly connected nodes often act as bridges across subgroups, causing structural centrality to merge otherwise distinct communities. Conversely, leaders identified through local density typically correspond to small, near-clique subgraphs that may merely reflect project-level co-participation (Menardi and De Stefano, 2022).

Here, we propose an attribute-driven perspective on leader identification. Leaders are characterised by high density in the attribute space. This means that they are surrounded by a concentration of nodes sharing similar attribute values, thereby serving as focal points around which communities form. This attribute-driven leader identification enables the detection of leaders who are not primarily structurally important, but rather attribute-representative. This enhances the interpretability of detected communities by anchoring them around actors whose attributes reflect the dominant features of their groups.

It is worth noting that, in principle, this formulation does not explicitly constrain leaders to be structurally connected to the nodes they represent, as density is computed purely in attribute space. In practice, however, this concern is mitigated by the nature of the attributes themselves. The approach assumes that selected node attributes exhibit homophily, meaning that attribute-dense regions tend to correspond to densely connected sets of nodes, making it unlikely that a structurally isolated node would emerge as a leader. A further discussion of this point, including potential extensions to settings where node attributes encode relational information, is provided in Section 4.5.

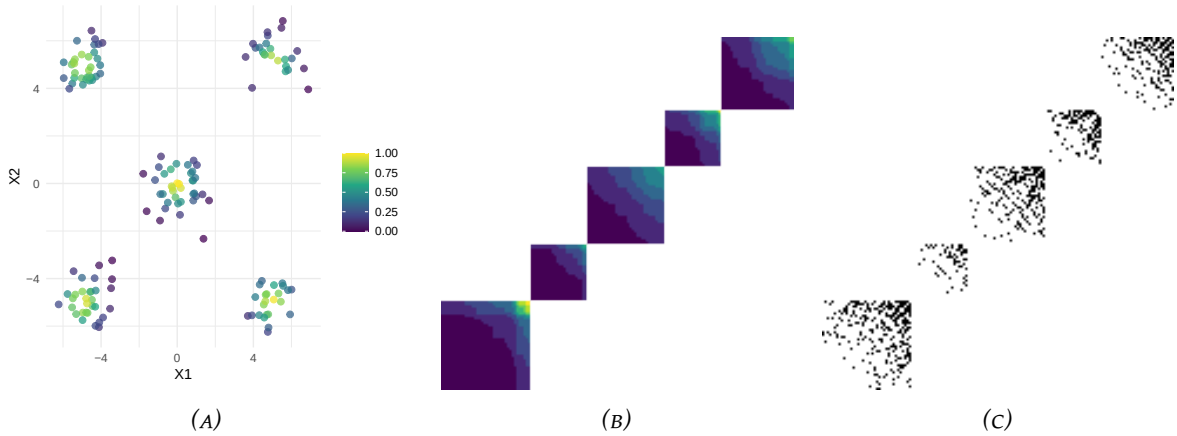


FIGURE 4.1: (a) Data points generated using a Gaussian model with 5 components, with colour representing density; (b) Block-diagonal edge probability matrix, with assortative-only blocks representing communities; (c) Block-diagonal adjacency matrix derived from (b).

Figure 4.1 illustrates the intuition. Panel (a) displays five well-separated clusters of data points, with colour representing local attribute density. Within each cloud, density varies, and only a few points (highlighted in yellow) lie at the centre of their group, acting as attribute-representative nodes. These nodes constitute ideal leaders in scenarios where similarity of attributes drives the emergence of communities. The central panel (b) translates this configuration into a homophily-consistent connectivity pattern, represented in a block-diagonal edge-probability matrix. Nodes that exhibit high attribute density are assigned higher connection probabilities to neighbouring nodes. Within each block, nodes are ordered by expected connectivity, with high-density leaders in yellow at the top of each diagonal block. Panel (c) shows the corresponding adjacency matrix, generated according to the probabilities in (b). The resulting network exhibits five assortative-only communities, each organised around its attribute-defined leader.

Although simplified for illustrative purposes, this example captures a structure in multiple attribute-coherent and degree-heterogeneous groups of nodes that we expect to observe in collaboration networks such as Horizon. Importantly, even in this idealised assortative setting, community recovery is non-trivial: within each block, most nodes are only sparsely connected, reflecting the heavy-tailed degree distributions typical of real-world data.

To uncover such structure, we extend the DeCoDe framework introduced by Menardi and De Stefano (2022). Let $G = (V, E, \boldsymbol{X})$ denote a node-attributed unweighted network, where $V = \{v_1, \ldots, v_i, \ldots, v_n\}$ is the set of nodes, $E = \{e_{ij}\}$ is the set of (undirected) edges, $i, j = 1, \ldots, n$, $i \neq j$, with $e_{ij} \in \{0, 1\}$ indicating the presence (1) or absence (0) of an edge between nodes v_i and v_j, ; $\boldsymbol{X} \in \mathbb{R}^{n \times p}$ is the attribute matrix whose i-th row $\boldsymbol{x}_i$ collects the p attributes associated with node v_i.

The DeCoDe approach leverages local density in the network space to identify

cluster cores and partition communities. Building on this idea, we integrate attribute information into the density estimation process. Specifically, we estimate the density for node v_i in a continuous attribute space,

$$\hat{\delta}(v_i) = \hat{f}(\boldsymbol{x}_i),$$

where $\hat{f}(\cdot)$ is a suitable probability density function estimator, reflecting the extent to which each node is representative of its neighbours in terms of relevant characteristics. We refer to the resulting method as the *Attribute-driven density for DeCoDe* (AttDeCoDe) approach.

Several alternative methods can be employed to estimate attribute-based densities. For clarity, we summarise the three approaches adopted in this work:

1. **k-Nearest Neighbours (kNN)** A node's density is computed as the inverse of the average distance to its k closest neighbours in the attribute space. This non-parametric estimator measures local attribute concentration without imposing any functional form on the underlying distribution.

2. **Gaussian Mixture Models (GMM)** GMMs (Bouveyron et al., 2019) provide a model-based density estimator by fitting a mixture of multivariate Gaussian distributions to the attribute space. Dense regions correspond to areas of high mixture likelihood. A node's density is calculated as the weighted sum of component likelihoods, where weights are the estimated mixing proportions.

3. **Component-specific multivariate Gaussian densities (GMM-based alternative)** Particularly suitable when the attribute space contains well-separated Gaussian-like clusters. This practical alternative first fits a GMM and then assigns each node to the component with the highest posterior probability. The density for the node is evaluated under that component's multivariate Gaussian distribution, defined by its estimated mean vector and covariance matrix. This procedure yields a collection of component-specific density estimates for each observation, according to its component membership, rather than a single global mixture density estimate.

These approaches offer a flexible way to capture homophily in continuous attributes and can be selected depending on the distribution and dimensionality of the attribute data.

4.2.2 Attribute-driven density-based community detection

The clustering problem in our context is framed in terms of detecting high-density regions on the network, where nodes are grouped according to a node-wise measure of density (Menardi and De Stefano, 2022). Two nodes belong to the same cluster if their densities exceed a given threshold and they are connected through a path in

G. This idea extends modal clustering to a relational setting: node-wise densities $\hat{\delta}(v_1), \ldots, \hat{\delta}(v_i), \ldots, \hat{\delta}(v_n)$ are estimated, and communities form by aggregating nodes around modal actors, i.e. local maxima of the density function.

In unweighted networks, for any threshold λ, the upper-level set $V_\lambda = \{v_i \in V : \hat{\delta}(v_i) \geq \lambda\}$ induces a subgraph $G_\lambda = (V_\lambda, E_\lambda)$, where $E_\lambda = \{e_{ij} \in E : v_i, v_j \in V_\lambda\}$. The connected components of G_λ represent density-based groups. As λ decreases, these components merge, producing a cluster tree that captures the hierarchical organisation of communities. The number of network clusters K is not fixed a priori, but is determined by the structure of this tree: clusters are identified as the connected components of G_λ at the lowest λ levels where they appear as leaves of the tree. Low-density nodes may either remain unassigned or be attached to nearby cluster cores. In addition to the detected number of clusters $\hat{K}$ and the cluster membership $C(v_i) \in \{1, \ldots, K\}$, the algorithm outputs an indicator $r(v_i)$ specifying whether node v_i acts as a cluster core ($r(v_i) = 1$) or as a non-core member ($r(v_i) = 0$).

Building on this framework, AttDeCoDe (Algorithm 2) generalises DeCoDe to node-attributed networks $G = (V, E, \boldsymbol{X})$, where node densities are computed from the attribute matrix $\boldsymbol{X}$ rather than from network structural properties. Attribute-based densities capture homophily and similarity in the attribute space, while the network topology still constrains the paths through which clusters form. This ensures that communities are both attribute-coherent and structurally feasible. AttDeCoDe therefore shifts the focus from purely structural density to attribute-informed density, allowing the clustering process to reflect meaningful variation in the attributes that characterise each node.

For completeness, we note that DeCoDe is recovered as a special case of AttDeCoDe. If the attribute-based density in Step 1 of Algorithm 2 is replaced with a structural density estimator based on G–such as degree, betweenness centrality, or other topology-derived measures–one obtains the original approach. DeCoDe is therefore a simpler variant that does not account for attribute information and relies solely on topology to define dense regions.

Nonetheless, estimating density from the feature matrix does not preclude structural information from influencing the identification of modal actors. In the context of networks constructed from bipartite data, such as organisation–project or author–paper networks, event-level attributes are transferred to nodes through projection. The resulting feature matrix encodes relational patterns: nodes participating in many events, or in events characterised by similar attribute profiles, naturally accumulate higher attribute density. Consequently, modal actors may emerge from a combination of attribute similarity and structural prominence, even though density is formally computed only from attributes. This mechanism is central to our application, where project-level textual descriptors shape each organisation's attribute representation and thus influence the density landscape from which modal actors are identified. More generally, alternative feature-construction strategies—such as node representations that explicitly combine structural prominence with attribute

information—could be adopted to encode both dimensions jointly; however, exploring these extensions lies beyond the scope of the present work.

4.3 Simulation study

4.3.1 Simulation design

To evaluate the empirical performance of the proposed method, we designed a simulation study based on a new generative framework that extends the DC-SBM. While the SBM is widely used to generate networks with community structures, the DC-SBM adds the flexibility to capture degree heterogeneity, thus the presence of leaders and less central nodes gravitating through them. Building on this foundation, our framework incorporates attribute-driven leader influence, allowing us to generate networks in which communities arise from both structural and attribute-based mechanisms.

Algorithm 2 AttDeCoDe

Input:
Attributed network $G = (V, E, \boldsymbol{X})$ with n nodes
Attribute-density estimator $\hat{f}(\cdot)$, as obtained from the approaches in Section 4.2.1

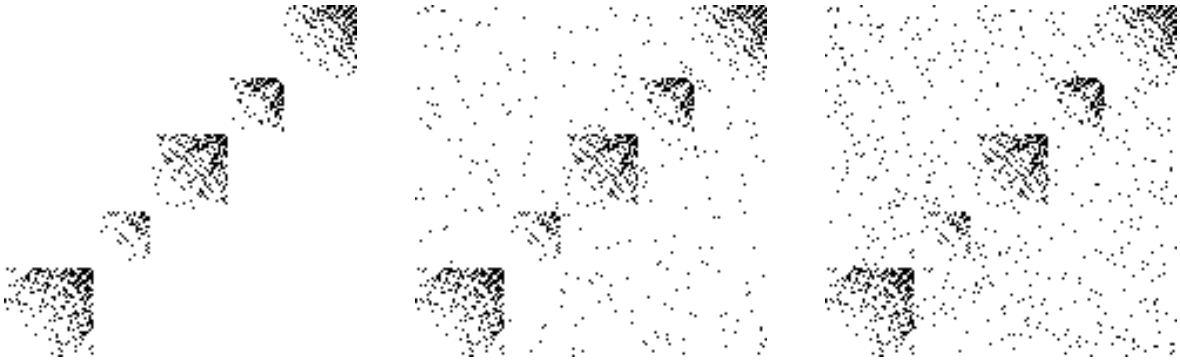
Step 1: Attribute-driven leader identification
Compute $\hat{\delta}(v_i) = \hat{f}(\boldsymbol{x}_i)$ where $\boldsymbol{x}_i$ is the attribute vector of node v_i

Step 2: Density-based community detection
for $0 < \lambda < \max_i \hat{\delta}(v_i)$ **do**
    Determine the upper level set:
    $V_\lambda = \{v_i \in V : \hat{\delta}(v_i) \geq \lambda\}$
    Build the induced subgraph:
    $G_\lambda = (V_\lambda, E_\lambda)$ where $E_\lambda = \{e_{ij} \in E : v_i, v_j \in V_\lambda\}$
    Identify the connected components of G_λ
end for

Build the cluster tree associating each λ with the number of connected components of G_λ

Identify clusters as the connected components of G_λ at the lowest λ levels (tree leaves)

Output:
Role $r(v_i) \in \{\text{core}, \text{member}\}$
Cluster membership $C(v_i) \in \{1, \ldots, K\}$
Number of detected clusters $\hat{K}$



(A) (B) (C)

FIGURE 4.2: Block-diagonal adjacency matrix representing a leader-based network partition in 5 communities; (a) assortative-only, 0% mixing; (b) 20% mixing; (c) 40% mixing.

The generative process proceeds in two stages. In the first stage, node-level attributes are sampled from a GMM using $M = \{5, 10\}$ mixture components. Different network sizes $n = \{50, 150, 300\}$ are considered. Community membership is determined according to a distribution, which is either uniform (yielding equal-sized communities) or drawn from a Dirichlet distribution (introducing cluster size imbalance). Each Gaussian component is defined by a fixed mean vector in $\mathbb{R}^2$ and a spherical covariance matrix ensuring well-separated clusters. Node densities are defined as their evaluation under the resulting Gaussian mixture density, so that higher density values correspond to nodes lying in more populated regions of the attribute space.

In the second stage, the network topology is generated using a DC-SBM with a community structure that matches the Gaussian data cluster structure, and with $K = M$. Within each community, nodes are connected with probability proportional to the product of densities δ_i. These δ values are derived from Gaussian densities and transformed to amplify differences between high- and low-degree nodes, thereby reproducing degree heterogeneity observed in real-world data. A detailed description of the construction of the δ values is provided in Appendix D.2.

Cross-community edges are introduced through a controlled mixing procedure, where a proportion of inter-community links are added. This mechanism is visible in the block-diagonal adjacency matrices displayed in Figure 4.2: from left to right, the originally ideal assortative structure becomes increasingly "blurred" as dark off-diagonal cells appear, producing a more realistic network topology.

Mixing is governed by a global parameter, μ, which regulates the extent to which nodes connect outside their own community. Increasing μ leads to a higher proportion of inter-community links and therefore to stronger overlapping between clusters. Namely, μ represents the ratio between a node's external links d_i^{ext} and its total degree d_i. Thus, when mixing is introduced, a fraction of μ links connects nodes in different communities (Lancichinetti et al., 2008).

To better reflect within-community preferential attachment, we allow mixing

rates to vary across nodes. Specifically, we assign each node a weight inversely proportional to node degree:

$$w_i = \frac{\overline{d}_{C(v_i)}}{d_i},$$

where $\overline{d}_{C(v_i)}$ is the mean degree in the community of node i. These weights are normalised to $[0, 1]$ and rescaled so that the expected mean mixing equals the target mixing value μ:

$$\mu_i = \min\left(w_i \cdot \frac{\mu}{\frac{1}{n}\sum_j w_j},\, 1\right),$$

with truncation at 1 to avoid invalid probabilities. The number of inter-community connections of each node is then given by $d_i^{ext} = \mu_i \cdot d_i$.

This design yields networks with realistic structural properties: communities exhibit heterogeneous density distributions driven by node attributes, while the level of inter-community mixing is regulated by $\mu = \{0, .2, .4\}$. By combining attribute-driven leader attraction with structural connectivity, our simulation framework provides a principled basis for assessing whether the proposed algorithm can recover the true partitioning of the network.

The proposed model is compared against several baseline methods. Specifically, we consider original DeCoDe, binary stochastic block model (SBM), Spectral Clustering (SC), Covariate-Assisted SC (CASC), and SC on Network-Adjusted Covariates (NAC). These approaches are commonly used in the literature and serve as representative competitors for both network-based and attribute-assisted clustering. The binary SBM is implemented via the `blockmodels` package (Leger et al., 2021a), while SC, CASC, and NAC are available in the `NAC` R package (Hu and Wang, 2023). For the SBM, the number of communities is selected by maximising the Integrated Classification Likelihood (ICL) (Côme and Latouche, 2015). We then use this ICL-optimal number of components for SC, CASC, and NAC. This choice is motivated by the close theoretical relationship between spectral methods and SBMs, established by Rohe et al. (2011). Relying on a single, statistically grounded selection rule also avoids the practical difficulties associated with method-specific alternatives: the cluster-number selection procedures available within the `NAC` package require applying each method across a range of values of K and identifying the optimal number via the elbow method, a criterion that is not always clearly detectable in practice and introduces an additional source of variability into the comparison.

The DeCoDe model provides a direct baseline for assessing the impact of our proposed extension. In our comparison, we evaluate AttDeCoDe under two main density specifications: (i) using the true density function employed to generate the synthetic networks, and (ii) using empirically estimated densities derived from the observed node attributes. For the empirical specification, we consider three alternative estimators. The first, referred to as the component-wise Gaussian density, computes separate densities for each Gaussian component identified by the fitted mixture model. The second estimator is the GMM density, obtained directly from the

mixture model implemented via the `mclust` package (Scrucca et al., 2023). Finally, we include a nonparametric baseline based on kNN density estimation with $k = 5$, implemented using the `FNN` package (Beygelzimer et al., 2024).

Clustering performance is assessed by measuring the agreement between the estimated and the true partitions, expressed through the Normalised Mutual Information (NMI) (Danon et al., 2005), which ranges from 0 (no agreement) to 1 (perfect agreement). We also compute the Adjusted Rand Index (ARI) (Hubert and Arabie, 1985b); however, simulation study results based on ARI are reported only in Appendix D, as they lead to conclusions fully consistent with those obtained using NMI and do not provide additional insights in this setting.

4.3.2 Simulation results

Among the empirical density estimators considered, we adopt the component-wise Gaussian density as the primary specification for our simulation study. This choice is motivated by the design of our generative model, where node attributes are drawn from distinct and well-separated Gaussian components. The GMM density tends to smooth across component boundaries, potentially obscuring inter-cluster separation, while the kNN estimator, though flexible, is less stable in low-dimensional settings with limited sample size. A detailed comparison of the performance of all three estimation approaches with the true density is provided in the Appendix (Figures D1).

Figure 4.3 shows the results of a study that investigates how network size and number of communities influence the ability of our AttDeCoDe approach to detect community structures in networks with equally sized communities. Results for networks with non-uniform community sizes are reported in the Appendix (Figure D2 and D3). Results suggest that our method performs well across a range of realistic network configurations. Because the DeCoDe approach is density-based, its performance is affected by the extent to which leaders form connections with high-density nodes in other communities. To evaluate this behaviour, we evaluate the performance of the detection methods applied to networks with varying proportions of inter-community links (mixing level μ).

Across scenarios for smaller networks ($n = 50$), both versions of the proposed method–AttDeCoDe using the true density (δ) and AttDeCoDe using the empirically estimated density ($\hat{\delta}$)–consistently outperform the grey and blue-colored baseline methods in terms of NMI. The improvement over the original DeCoDe is most pronounced when mixing is introduced, where attribute information substantially enhances community separability. As the mixing increases, all methods show a decline in performance due to reduced structural distinctiveness among communities, but AttDeCoDe maintains the highest stability across the full range of mixing.

Increasing network size improves performance for baseline algorithms, indicating better recovery of both structural and attribute-based patterns in larger samples. In contrast, AttDeCode performance remains stable across network sizes. The effect of

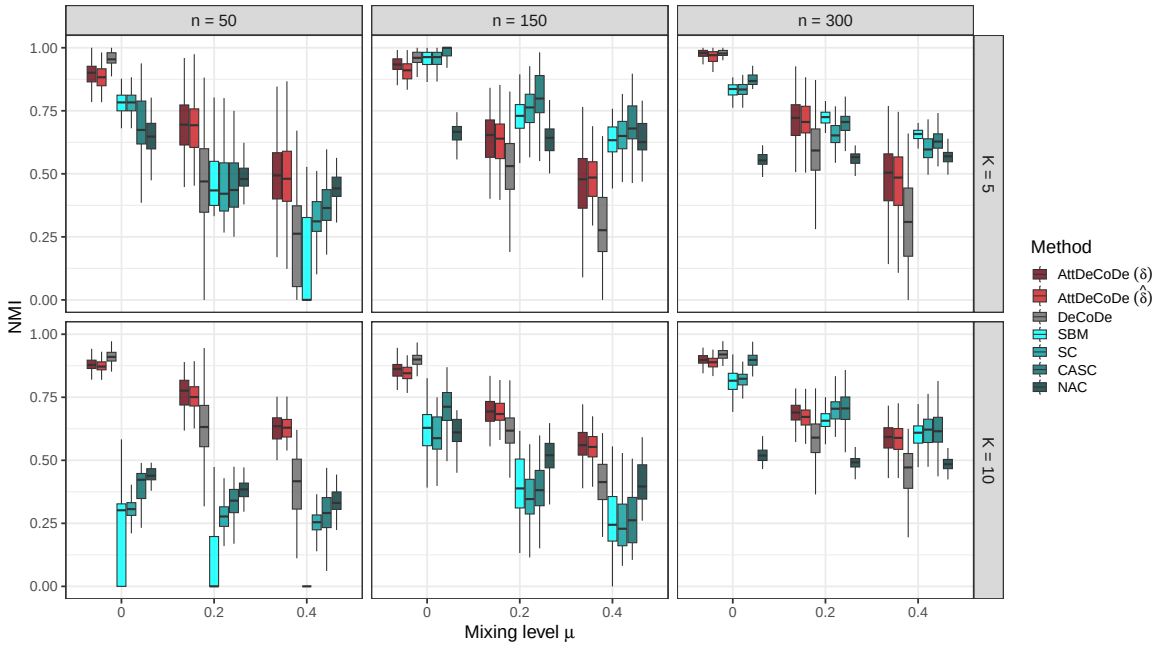


FIGURE 4.3: Normalised Mutual Information (NMI) distribution across different network sizes and numbers of communities (K) for uniform community sizes. Results are reported for all competing methods, including AttDeCoDe, DeCoDe, binary stochastic block model (SBM), Spectral Clustering (SC), Covariate-Assisted SC (CASC), and SC on Network-Adjusted Covariates (NAC).

the number of components in the attribute space is also visible: for $M = 10$, the nodes are grouped in smaller communities, and the network is more fragmented, especially for smaller networks, making community detection more challenging. In this setting, the gain from explicitly modelling density heterogeneity (as in AttDeCoDe) remains evident, whereas simpler structural approaches exhibit a decline in performance.

Overall, these results confirm that AttDeCoDe effectively integrates attribute-driven density information, yielding robust community recovery even under moderate inter-community mixing.

4.4 Application

4.4.1 The Les Misérables network

The evaluation of the detection performance involves examining how our method performs on a popular dataset, where community membership and attributes of nodes are available. The popular Les Misérables character network (Figure 4.4) describes the interactions between 77 characters in Victor Hugo's novel Les Misérables (Knuth, 1993). The edges represent the co-appearance of characters in one or more chapters

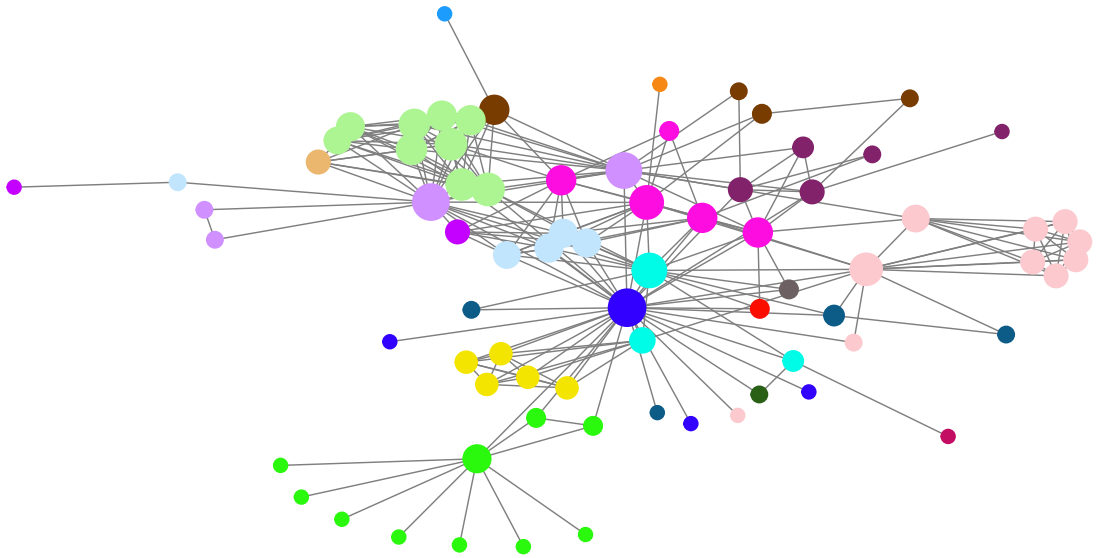


FIGURE 4.4: Les Miserables network partition in 20 communities with node size representing degree and node colour representing community membership.

of the novel. The ground truth membership corresponds to the early appearance of each character in the book, resulting in $K = 20$ small communities. We used textual descriptions of characters as node attributes, which are accessible from the dataset provided in (Kalugin, 2015). Similarly to how we handled the data in Chapter 3, we transformed the character descriptions into a tf–idf (term frequency-inverse document frequency) matrix. This numerical representation allows us to incorporate the textual information on the characters into our analysis, enabling a richer understanding of the network.

For estimating the attribute-based density, we fit a GMM with a large number of components, namely $M = n/2$. This approach is supported by theoretical findings that a GMM can approximate any smooth density function with arbitrary accuracy, given a sufficient number of components (Goodfellow et al., 2016; Nguyen et al., 2020). Additionally, this choice is preferable for this dataset, as the small number of data points would lead to poor estimates using methods like kNN. To visualise the attributive information, we applied multidimensional scaling (MDS) to the cosine distance matrix computed from the tf–idf document-term matrix. In the resulting two-dimensional embedding (Figure 4.5a), nodes with higher GMM density values are represented with lighter colours (i.e., yellow and green) and tend to appear in the denser regions of the plot. The distribution of the GMM density values differs from that of the node degree or the local density used in the original Decode paper; all quantities are normalised to ensure comparability. As the degree, the GMM density distribution exhibits a peak at low values, but it also shows a secondary local peak around 0.5 (Figure 4.5b).

Table 4.2 presents clustering performance across different methods. Among

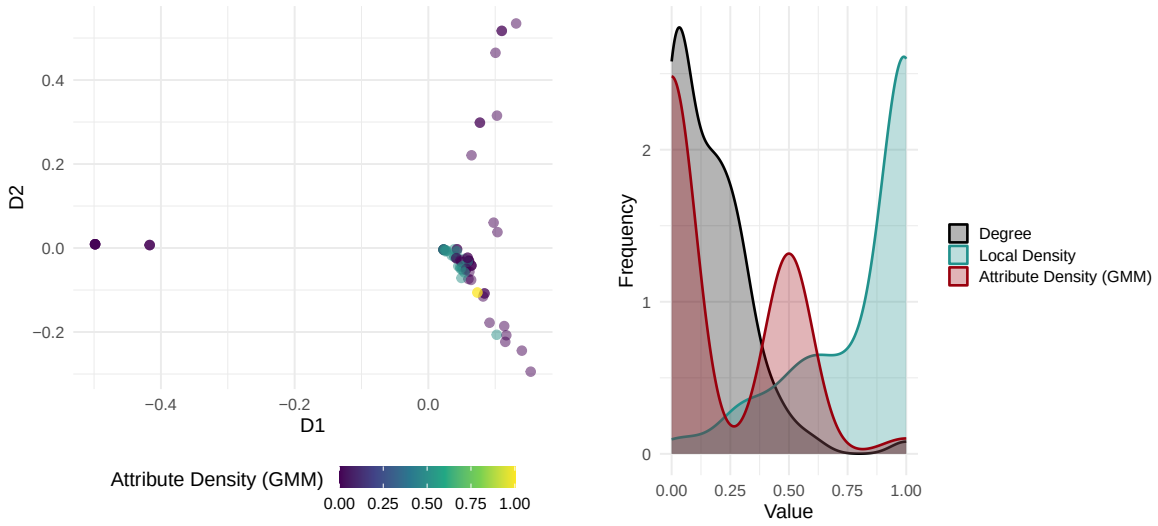


FIGURE 4.5: Les Misérables network. (a) Two-dimensional MDS projection of node attributes with lighter colours indicating higher GMM density; (b) normalised distributions of GMM density, node degree, and local density.

the DeCoDe variants, using node degree alone performs poorly, failing to recover meaningful community structure (NMI = 0, ARI = 0, $\hat{K} = 1$). Incorporating local density markedly improves performance (NMI = 0.76, ARI = 0.40, $\hat{K} = 35$), reflecting the benefit of considering neighbourhood density for identifying communities. The proposed method, AttDeCoDe, recovers 22 clusters, which is close to the true number, indicating a smaller degree of over-segmentation than the local density variant. AttDeCoDe achieves the highest agreement with the ground truth among all tested methods, with NMI = 0.78 and ARI = 0.50, outperforming other standard methods such as SBM, SC, CASC and NAC. This suggests that incorporating attribute-driven density information into the community detection process improves the recovery of the underlying community structure.

4.4.2 The hydrogen Horizon network

This section applies the proposed AttDeCoDe framework to the hydrogen Horizon collaboration network, a large-scale empirical setting in which community formation is driven not only by network structure but also by thematic similarity across organisations.

Topic-specific clouds tend to form around organisations that dominate particular research areas, whereas actors engaged in more interdisciplinary, emerging, or weakly connected topics may occupy a more dispersed position within the network. Our approach estimates attribute-based node density to identify thematic focal points and then applies DeCoDe to delineate the subgraphs that form around them. This

perspective is consistent with the logic of Horizon collaborations: organisations engaged in similar research areas tend to cluster around shared priorities, while differences in expertise, mission, or sector generate boundaries that limit the formation of ties.

To identify clusters driven by homophilous influence, we construct node-level attributes by assigning each organisation the complete set of keywords attached to the hydrogen projects in which it participates. This yields a standardised thematic profile for each organisation, allowing us to assess whether proximity in scientific domains is reflected in the observed clustering patterns. By grounding community detection in project-level attributes, our framework identifies leaders as organisations that are both thematically central–exhibiting high attribute density–and structurally well positioned to attract collaborations across multiple projects, thereby accumulating a broad and diverse keyword profile.

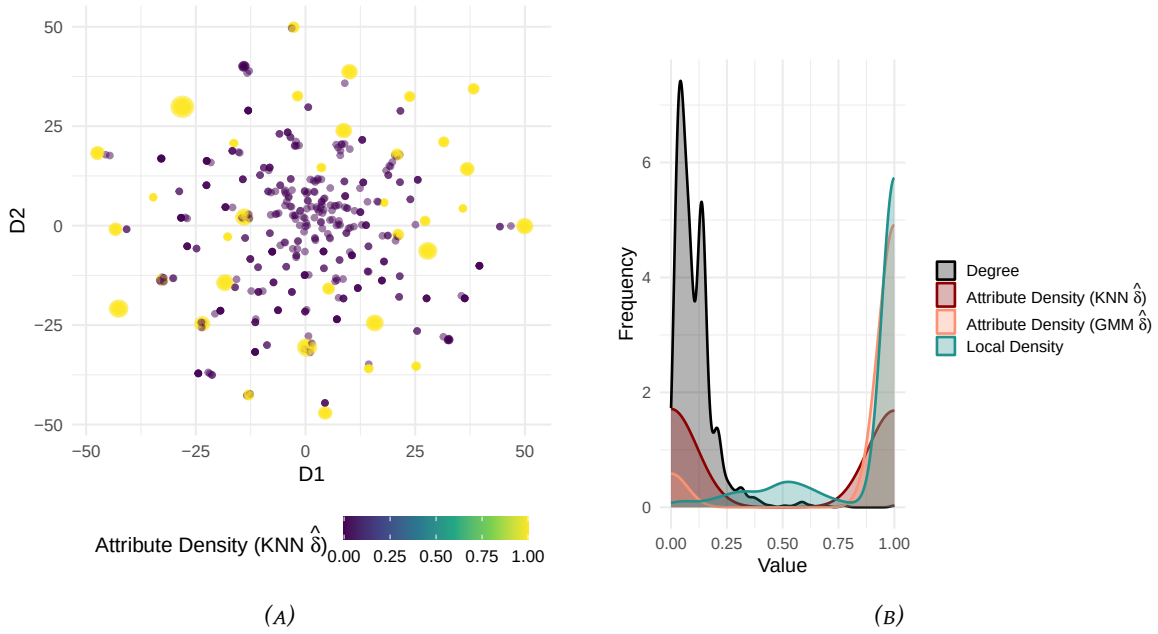
We represent each organisation through a dense semantic embedding of its associated keywords and estimate local concentration in the resulting numerical attribute space. Specifically, we generate sentence embeddings from the concatenated keyword lists using the all-MiniLM-L6-v2 SentenceTransformer model (Hugging Face, 2024), which yields 384-dimensional vectors in a dense semantic space. This embedding approach allows us to capture higher-order semantic relationships among keywords–beyond simple term co-occurrence, and to represent organisations in a continuous attribute space suitable for density estimation.

To visualise this high-dimensional attribute data, we apply t-SNE to the embeddings. t-SNE (van der Maaten and Hinton, 2008) is widely used for dimensionality reduction of embedding spaces produced by language models, as it preserves local neighbourhood structure and facilitates the visualisation of underlying clustering patterns. In the resulting two-dimensional embedding space (Figure 4.6a), attribute-defined dense regions appear as yellow points close in the space.

For estimating the attribute-based density, we fit a GMM with a relatively large

Method	NMI	ARI	$\hat{K}$
DeCoDe (Degree)	0.00	0.00	1
DeCoDe (Local Density)	0.76	0.40	35
AttDeCoDe	**0.78**	0.50	**22**
SBM	0.20	0.54	6
SC	0.32	0.63	6
CASC	0.02	0.30	6
NAC	0.24	0.54	6

TABLE 4.2: Les Misérables network. Agreement in terms of Normalised Mutual Information (NMI) and Adjusted Rand Index (ARI) between estimated and early appearance partition, along with estimated number of clusters ($\hat{K}$), across community detection methods.



(A) *(B)*

FIGURE 4.6: Distribution of node density in the hydrogen Horizon network (a) Two-dimensional TSNE projection of node attributes with lighter colours indicating higher kNN density; (b) normalised distributions of kNN density, node degree, and local density.

number of components (M). Given the high number of points (n), using $n/2$ components–as in the benchmark network example–is computationally infeasible, as the Expectation–Maximization algorithm in `mclust` fails to converge due to overparameterization. Instead, we constrained M to be greater than 100 and allowed the optimal number of components to be automatically selected by the Bayesian Information Criterion (BIC).

We also considered a non-parametric kNN density estimator with $k = 5$. This relatively small value of k allows for capturing local variations in attribute concentration while maintaining robustness against noise.

Both GMM and kNN attribute densities capture similar general patterns, being able to identify these visually dense clouds. However, their shapes differ (Figure 4.6b). The GMM estimator assigns the maximum density to approximately 89% of the nodes, collapsing most of the attribute space into a single high-density mode and failing to distinguish a substantial proportion of low-density points. By contrast, the kNN density assigns high density to around half of the nodes and spreads the remaining ones across lower values, thereby successfully identifying both the localised dense regions visible in the two-dimensional space, as displayed in Figure 4.6a.

This partition between low- and high-density nodes is not easily captured by structural centrality measures such as degree or local density, which show unimodal distributions–degree concentrated at low values and local density concentrated at

high values–providing no clear way to identify meaningful high-density regions. These divergent behaviours also complicate the choice of a centrality measure in the original DeCoDe framework. At the same time, the heterogeneous degree distribution of the network, characterised by hubs that span multiple communities, reinforces the need for a density-based approach. Applied to the Horizon data, DeCoDe using degree produces a trivial solution with a single cluster, offering no analytical insight, whereas DeCoDe using local density heavily over-segments the space, partitioning the 825 organisations into 115 clusters with an average size of only seven nodes.

Instead of relying on structural prominence alone, modelling density in the attribute space provides an appropriate means of uncovering community structure driven by thematic proximity. This is particularly relevant in the Horizon network, where the attractiveness of cluster representatives–and the formation of groups around them–is plausibly driven by similar research interests rather than by purely structural prominence.

When AttDeCoDe is applied using the GMM-based density, the method yields 11 clusters, dominated by one large component of 767 organisations and a set of very small, highly isolated groups–often behaving as near-cliques. Among these, only two of the ten small clusters contain connected organisations participating in at least two distinct projects. These clusters exhibit thematic coherence (e.g., hydrogen materials, electrolysis, mobility applications) and consistent institutional and geographic profiles, suggesting that GMM detects only the most concentrated local maxima in the attribute space. However, its inability to distinguish the broader low-density regions of the space limits its interpretability. In contrast, AttDeCoDe using the kNN-based attribute density (Figure 4.6) provides a more balanced and informative representation, partitioning the organisations into 52 dense communities (Figure 4.7 and 4.8).

Characterisation of the detected clusters

The clusters, obtained using AttDeCoDe with kNN, display a highly uneven size distribution, ranging from small groups of 3–5 organisations to large clusters comprising more than 60 members. The majority of clusters are moderately sized (10–30 organisations), reflecting the coexistence of compact, cohesive subgroups and broader, cross-sectoral alliances typical of Horizon consortia. This heterogeneity in size suggests that community formation in the Horizon network operates at multiple scales, with both tightly focused collaborations and large, integrative research domains coexisting within the same system.

The correspondence between cluster size and project participation further illustrates this diversity. While nine clusters include organisations participating in a single project only, a small number of large clusters aggregate a substantial number of projects–up to more than 50 in the largest case (red cluster 36 in Figure 4.7). This pattern indicates the presence of broad research domains that integrate multiple collaborative initiatives and act as structural backbones of the network.

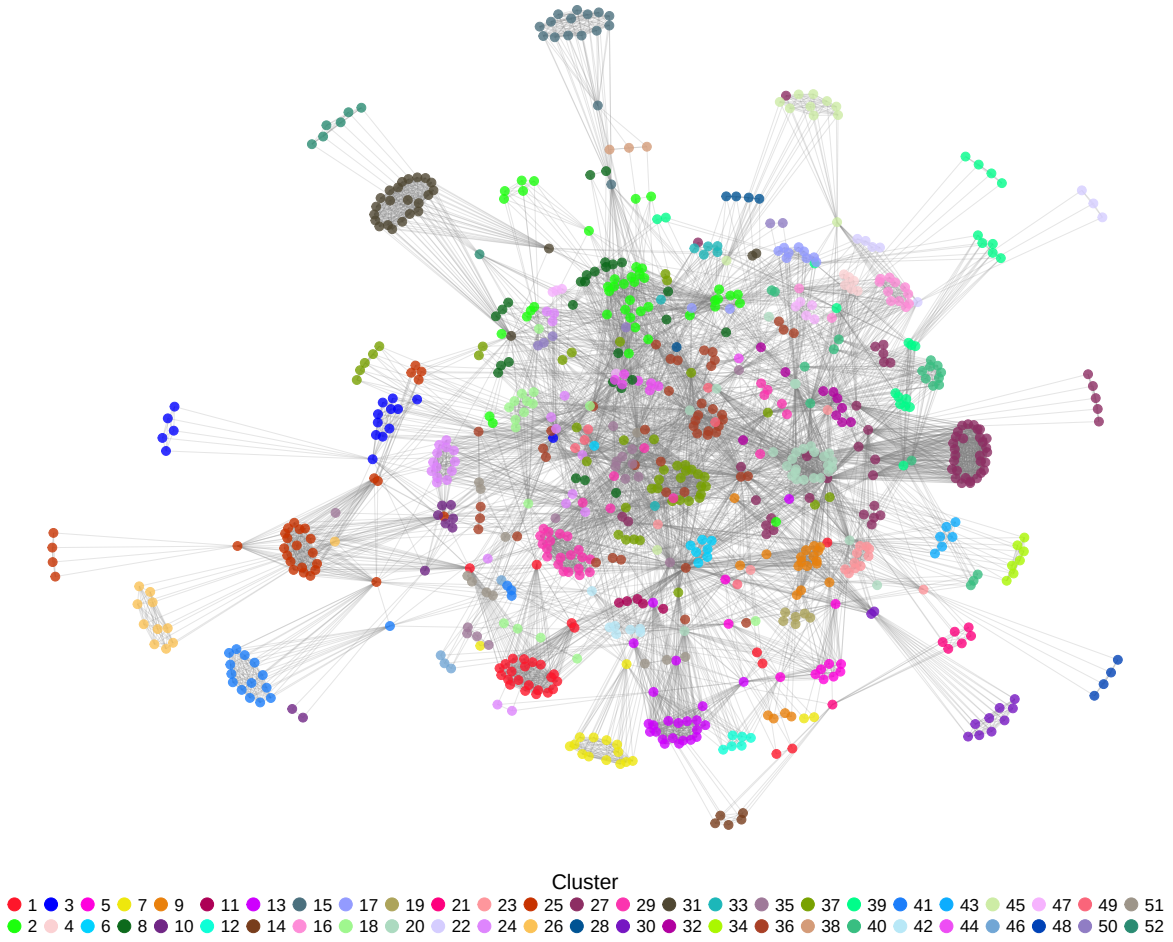


Figure 4.7: Hydrogen Horizon network partition in communities using AttDeCoDe with kNN density on the project keyword embeddings.

Tables 4.3 and 4.4 provide a detailed quantitative characterisation of the clusters ranked among the top five by size (allowing for equal frequencies) in terms of geographical composition, activity type, and dominant research topics. The composition of the clusters is consistent with the overall structure of the dataset and with the design principles of Horizon consortia. Geographically, these clusters are generally transnational, involving organisations from several highly represented countries–most notably France, Italy, Germany, and Spain. These results are consistent with Balland et al. (2019b), who show that EU Framework Programmes foster highly integrated, transnational research networks, with certain organisations acting as key connectors between countries, reflecting patterns of collaboration similar to those observed in our clusters. However, the degree of concentration varies in other identified clusters. For example, cluster 25 is largely national, being dominated by Greek organisations (20 out of 31), whereas clusters 27, 36, and 37 display a more balanced international composition. This pattern indicates that some research areas remain embedded in

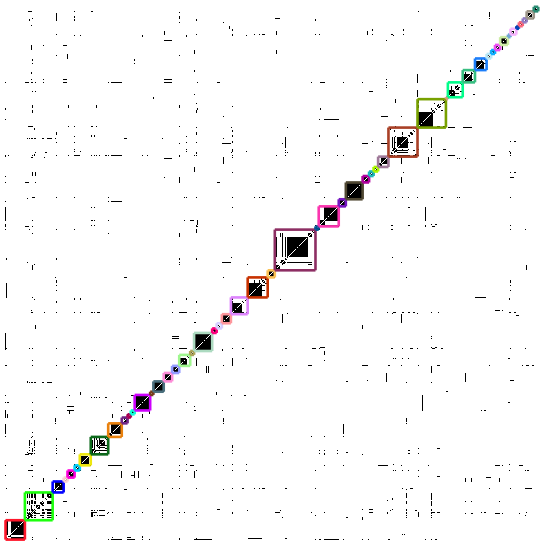


FIGURE 4.8: Block-diagonal adjacency matrix of the hydrogen Horizon network, representing intra- and inter-community links identified by AttDeCoDe.

national ecosystems, while others operate at a European scale.

In terms of activity type, large clusters are predominantly composed of private research companies, reflecting the applied orientation of Horizon projects, but they

TABLE 4.3: Composition of the five largest clusters (allowing for equal frequencies) by country and activity type.

Cluster	Cluster size	Main countries (N)	Activity types (N)
27	63	FI (17), RoE (16), DK (8), DE (6)	PRC (33), HES (10), RoE (10), REC (8)
36	45	DE (8), FR (8), BE (5), ES (5), IT (5)	PRC (32), HES (8), REC (4)
37	44	ES (12), DE (8), FR (5), IT (5), NL (4)	PRC (30), REC (7), HES (6)
2	43	ES (9), RoE (8), DE (7), IT (6), NL (4)	PRC (29), HES (10), REC (4)
29	31	OTH (15), IT (13), BE (2)	PRC (22), REC (5), HES (3)
25	31	EL (20), RoE (3), NL (2), UK (2)	PRC (19), REC (5), HES (3), PUB (2)

TABLE 4.4: Composition of the five largest clusters (allowing for equal frequencies) by dominant research topics. Project keywords associated with cluster leaders in bold.

Cluster	Top 5 keywords (N)
27	**natural gas** (43); electrolysis (25); fuel cells (17); catalysis (8); machine learning (8); synthetic fuels (8)
36	**electrolysis** (33); **geometry** (21); **productivity** (21); catalysis (12); fuel cells (12); mining and mineral processing (12)
37	**metallurgy** (25); **thermodynamic engineering** (25); electrolysis (17); fuel cells (14); coating and films (8)
2	**catalysis** (27); **aliphatic compounds** (20); **lipids** (20); electrolysis (11); coating and films (10)
29	**ecosystems** (25); **fuel cells** (25); **sustainable economy** (25); natural gas (8); sensors (8)
25	**public transport** (23); **digital electronics** (23); sustainable economy (7); electrolysis (7); coating and films (6)

systematically include research organisations and higher-education institutions. This cross-sectoral mix aligns with Horizon programme rules, which promote partnerships across countries and activity types, particularly between academia and industry. The findings are in line with previous studies on EU-funded collaboration networks, which document spatial heterogeneity in community formation (Barber and Scherngell, 2013; Morea et al., 2024).

The dominant topics reported in Table 4.3 further characterise the largest clusters. For each cluster, the table reports the most frequent project keywords (top five by rank, allowing for equal frequencies), with those associated with the cluster leader highlighted in bold. Leaders correspond to the cluster cores identified by AttDeCoDe and represent organisations located in high-density regions of the attribute space, around which the remaining cluster members are organised.

Across all large clusters, fuel cells and electrolysis emerge as recurrent transversal themes, reflecting the overarching focus of the funded research and the strategic priorities of European hydrogen policy. At the same time, each cluster is structured around a distinct combination of secondary topics, revealing meaningful thematic differentiation.

Cluster 36 combines *electrolysis* with *productivity*, *geometry*, and mining-related topics, suggesting a focus on large-scale industrial processes and optimisation. Cluster 37 is centred on *metallurgy* and *thermodynamic engineering*, with a clear emphasis on applied engineering and materials science. Cluster 2 exhibits stronger links to *catalysis*, organic compounds, and chemical processes, while cluster 29 integrates sustainability-oriented themes such as *ecosystems* and *sensors* with energy-related topics. Finally, cluster 25 is characterised by *public transport* and *digital electronics*, pointing to application domains related to mobility and infrastructure rather than to core hydrogen technologies alone.

This pattern–shared core themes coupled with differentiated thematic specialisation–is consistent with previous studies of EU-funded collaboration networks, which document the emergence of field-specific communities organised around common strategic priorities but diversified along technological and application-oriented dimensions (Barber and Scherngell, 2013; Balland et al., 2019a).

Beyond the large clusters, only 13 clusters contain topics that are unique to them. Among these, two clusters–37 and 52–are particularly distinctive, as in both cases organisations participate in projects for which more than half of the associated keywords are unique to that cluster. Cluster 37 exhibits a strong technological and industrial orientation, combining *metallurgy*, *thermodynamic engineering*, *fuel cells*, and advanced manufacturing. This focus is directly reflected in the leader's keyword profile, reinforcing the interpretation of this cluster as a cohesive engineering-oriented community. In contrast, cluster 52 is markedly research-driven, focusing on *tissue engineering*, *stem cells*, *piezoelectrics*, and chemical engineering. Its members are primarily academic and research institutions, mostly located in France, Belgium, and other associated countries. The leader of this cluster exhibits a broad yet coherent thematic profile–spanning *chemical engineering*, *coating and films*, *electrolysis*,

photocatalysis, *piezoelectrics*, *solar energy*, *stem cells*, and *tissue engineering*–which captures the interdisciplinary nature of the community at the intersection of life sciences, materials science, and energy-related technologies. Importantly, such a niche structure is not apparent from structural connectivity alone and emerges clearly only when attribute-based density is taken into account.

Overall, these results show that the proposed approach is able to recover communities that are not only structurally cohesive but also thematically interpretable. The identified leaders provide a clear thematic anchor for each community, facilitating interpretation of the underlying research focus. Large clusters correspond to broad, multidisciplinary research domains that integrate multiple projects, countries, and sectors, whereas smaller clusters capture more specialised thematic niches. The joint analysis of cluster size, organisational composition, thematic focus, and leader profiles thus offers a comprehensive view of how collaboration patterns in the Horizon network emerge from the interplay between structural connectivity and attribute-driven similarity.

Thanks to the proposed approach, it emerges that EU-funded hydrogen research collaborations form a multi-scale, cross-sectoral, and largely transnational landscape. Core themes such as *fuel cells* and *electrolysis* recur across clusters, while secondary topics differ between communities, reflecting both thematic specialisation and the strategic priorities of Horizon consortia. Within each cluster, identified leader organisations serve as thematic anchors around which other members are structured, ensuring that the network combines cohesion with specialised focus and captures the coexistence of European-scale integration and local research ecosystems.

4.5 Final remarks

This paper has proposed an attribute-driven extension of the DeCoDe framework of Menardi and De Stefano (2022), designed for settings in which group structure is shaped jointly by network topology and node attributes. The central methodological contribution is to shift the identification of leaders from purely structural prominence to high density in an attribute space, thereby allowing communities to be organised around nodes that are representative in terms of characteristics.

A complementary simulation design extends the DC-SBM (Karrer and Newman, 2011) by combining block structure with attribute-driven degree heterogeneity, providing a principled environment to evaluate our novel approach against established alternatives. Across a wide range of sparse network configurations, incorporating attribute-driven density in the leader-identification step improves the community recovery relative to both purely structural and attribute-assisted competitors. Notably, the DC-SBM is not included among the baseline competitors in the current comparison; nevertheless, a systematic comparison with DC-SBM represents a natural and valuable direction for future work, particularly in settings characterised by a strong degree heterogeneity.

The empirical application further demonstrates the practical value of the method. In the Horizon collaboration network, AttDeCoDe uncovers thematically coherent communities and succeeds in distinguishing small, specialised thematic clusters from broader cross-sectoral groupings–patterns that are not easily identifiable when relying on structural centrality alone. The proposed method also allows to discover (thematic) leader organisations that guide the observed community structure and play a crucial role in establishing and consolidating research fields. Furthermore, the identification of such actors may be crucial for funding distribution policies (Morea et al., 2024).

Although AttDeCoDe computes density exclusively from node attributes, structural information may still influence which nodes emerge as modal actors. This is particularly evident in our bipartite setting, where event-level information projected onto nodes implicitly encodes participation structure: organisations involved in many events, or in events with similar thematic profiles, accumulate attributes that place them in denser regions of the feature space. As a result, the attribute representation captures aspects of structural prominence.

More generally, this mechanism arises when attributes carry relational content, such as interaction weights or other attribute-based measures of collaboration intensity. For example, quantities such as the total monetary value of the projects in which an organisation participates (Morea et al., 2024) could likewise be incorporated as node-level density indicators by computing node strengths—defined as the sum of monetary values of the project collaborations incident to each organisation—thereby providing an alternative representation of leadership in collaborative settings. Likewise, node embeddings obtained through representation learning techniques–whether based on random walks, spectral methods, or graph neural networks–can be constructed by fusing structural and attributive information (Cai et al., 2018; Cui et al., 2018). These examples illustrate natural extensions of the AttDeCoDe framework. More generally, they highlight a key advantage of AttDeCoDe: although density is formally defined in the attribute space, the method remains sensitive to structural variation whenever relational information is directly or indirectly embedded in the node attributes.

Beyond these examples, the proposed methodology opens several additional avenues for further research. First, our implementation assumes a static, unweighted network. Extending AttDeCoDe to weighted or temporal networks (Balland et al., 2019b; Morea et al., 2024) would broaden its applicability, enabling a better understanding of how communities evolve as collaborations expand, consolidate, or reorganise over time. While the DeCoDe framework can accommodate weighted networks, doing so is non-trivial: empirical weight distributions are typically highly skewed and reflect factors such as consortium size, visibility, or resource availability, rather than purely structural interaction.

Second, the method inherits sensitivity to the choice of the density estimator (e.g. the number of mixture components in GMMs or the choice of k in kNN) and in the construction of high-dimensional attributes (such as sentence embeddings for textual data). A systematic study of tuning strategies, uncertainty quantification, and

robustness diagnostics would therefore be valuable for users.

Third, while our simulation design relies solely on GMM-simulated attributes, this choice reflects both the motivating application and the common use of GMMs to model textual embeddings (Clinchant and Perronnin, 2013). Nonetheless, the framework can incorporate alternative mixture models for which well-established density estimation theory exists, such as t-mixture models McLachlan and Peel (2005), allowing the exploration of settings with heavier tails or more heterogeneous attribute distributions.

Lastly, the presented empirical application focuses on a relatively narrow research domain relevant to innovation dynamics and relies exclusively on the CORDIS database. This limits the scope to EU-funded initiatives and may overlook collaborations supported by national programmes, regional authorities, or private investment. Integrating information from alternative funding sources would offer a broader view of the research ecosystem, though comparable datasets are not currently accessible.

Despite these caveats, the proposed framework shows that attributive information can be integrated into DeCoDe in a computationally tractable manner, detecting clusters that are both structurally cohesive and thematically interpretable. We anticipate that AttDeCoDe will be useful in a range of applications where community formation is driven by the joint influence of network structure and node characteristics, such as scientific collaboration, political alliances, and organisational ecosystems, and that it will serve as a starting point for further methodological developments at the interface of network statistics and multivariate density estimation.

Part III

Community-level core-periphery patterns

Introduction

The increasing collaboration and interdependence among researchers reflects the need for multidisciplinary approaches to address social, political, economic, and technological challenges (Wuchty et al., 2007). Uncovering structures in collaborative research networks is crucial for understanding how ideas spread, and capabilities evolve, ultimately driving scientific productivity and the generation of new knowledge (De Stefano et al., 2011). To deepen understanding of these complex mechanisms, researchers have developed statistical approaches aimed at identifying structures in collaboration networks, with large-scale co-authorship network analyses enabled by the increasing availability of bibliometric data (Donthu et al., 2021).

A growing body of work focuses on meso-scale network structures, with community detection being a leading theme. In research and innovation systems, communities are frequently shaped around influential actors, but also thematic areas or regional collaborations, forming social circles (Alba and Moore, 1978; Alba and Kadushin, 1976). Acknowledging the interdependencies among community members is crucial for understanding how collaboration evolves within the scientific community. Equally important is to examine how community structures interact with other organisational patterns, such as disassortativity, core-periphery dynamics, or the presence of hubs (Fortunato, 2010; Legramanti et al., 2022; Kojaku and Masuda, 2018).

While community structures in collaboration networks have been extensively studied (Newman and Girvan, 2004; Lužar et al., 2014; Menardi and De Stefano, 2022), relatively less attention has been paid to their arrangement in core-periphery structures (Zelnio, 2012; Cugmas et al., 2015; Karlovčec et al., 2016; Sedita et al., 2020; Wedell et al., 2022). This structure is used to describe systems where a cohesive, central group, the *core*, interacts densely within itself, while a more loosely connected group, the *periphery*, maintains ties mainly with the core (Yanchenko and Sengupta, 2023; Tang et al., 2019).

Despite the potential interplay between community and core-periphery structures, their combined presence in social networks has rarely been explored (Legramanti et al., 2022). An initial strategy to address this gap is to consider methods capable of capturing more general block structures in network data. Notably, SBMs provide flexible frameworks for uncovering core-periphery and hierarchical structures (Gallagher et al., 2021; Yanchenko and Sengupta, 2023).

Building on this idea, we develop a core-periphery classification framework specifically designed to detect nested meso-scale structures. Unlike traditional methods that identify core-periphery organisation at the node level, we extend the method to identify core and peripheral groups of nodes. Specifically, assuming a

community structure in the network, we focus on a binary partition of the communities, inspired by the two-block model of Borgatti and Everett (2000). To the best of our knowledge, this is the first approach to explicitly identify nested core-periphery organisation in collaboration networks, revealing how such structures emerge between communities rather than, or as well as, within them. In doing so, we introduce a novel community-level framework that redefines the focus of core-periphery classification beyond individual nodes.

In Section 5.1 we present the motivating collaboration network that inspired our study, along with the associated research questions. Section 5.2 provides the background and methodology: we introduce the new community-level core-periphery network structure under investigation (Section 5.2.1) and, in Section 5.2.2, we formally define the community-based core-periphery classification problem and describe our proposed approach. Simulation studies in Section 5.3.1 assess the performance of the proposed approach. Section 5.4 discusses the findings of the collaboration network analysis. Finally, Section 5.5 concludes with a discussion and directions for future research.

Relevant literature

The concept of *core-periphery* has its roots in economic and sociopolitical theories (e.g. Prebisch, 1949; Galtung, 1971) and was formalised in network science through blockmodeling approaches in the 1970s (Breiger et al., 1975; Mullins et al., 1977; White et al., 1976). Since then, it has become a central idea for describing heterogeneity in networks, with applications spanning the social, economic, and biological sciences (Hidalgo et al., 2007; Rombach et al., 2012; Bassett et al., 2013). However, as recent surveys emphasise (Tang et al., 2019; Yanchenko and Sengupta, 2023), there is no single, universally accepted definition of what constitutes a core or a periphery. Instead, multiple conceptualisations coexist, each relying on different assumptions about how core and peripheral nodes should relate to one another. As a consequence, researchers are often faced with methods that–while all labelled "core-periphery detection"–encode distinct and sometimes incompatible visions of the concept.

Two main definitions, each derived from exemplary models, dominate the literature. The first is associated with the layered (or k-core) model, which interprets coreness as a gradual property rather than a strict dichotomy. In this framework, approaches for core-periphery detection embed nodes in nested layers, with inner layers representing the most central core (Batagelj and Zaveršnik, 2003; Goltsev et al., 2006; Hébert-Dufresne et al., 2016; Gallagher et al., 2021). These approaches, also referred to as transport-based methods, exploit flow dynamics or path structures, typically relying on geodesic distances or random walks to assign coreness scores (Lee et al., 2014; Cucuringu et al., 2016). They allow for multiple levels of peripheral integration and have been widely applied to capture hierarchical organisation in large-scale networks.

The second major paradigm corresponds to the two-block (hub-and-spoke) model, formalised by Borgatti and Everett (2000). In this formulation, core nodes are densely connected both among themselves and with peripheral nodes, whereas peripheral nodes maintain only sparse interconnections. This hub-and-spoke view assumes overall network connectivity, with the periphery integrated through its ties to the core. The Borgatti and Everett (2000) framework remains one of the most widely adopted approaches to core-periphery detection in practice, despite subsequent developments and alternative formulations (Boyd et al., 2006; Brusco, 2011; Lip, 2011; Zhang et al., 2015; Gallagher et al., 2021; Estévez and Nordlund, 2025). A key strength of this family of models is its simplicity: nodes are dichotomised into either core or periphery, enabling straightforward interpretation. Moreover, the two-block model has close connections with the SBM (Karrer and Newman, 2011), as both evaluate connectivity patterns against an idealised density-based structure.

Despite its popularity, the Borgatti and Everett (2000) framework presents several limitations when applied to real-world networks. Kojaku and Masuda (2018) argued that simple hub-and-spoke structures lack the flexibility to account for interesting patterns other than those already explained by degree heterogeneity. Instead, they suggest that identifying a meaningful core-periphery organisation requires identifying additional substructures within the same network, such as communities, bipartite patterns, or overlapping core-periphery pairs. Their work specifically focuses on uncovering the latter (Kojaku and Masuda, 2017, 2018). Other studies have examined similar multi-core structures, where several dense, mutually interacting cores coexist alongside peripheral regions Rombach et al. (2012); Zhang et al. (2015); Cugmas et al. (2015). Such a multi-core structure arises naturally in SBM formulations, particularly in their hierarchical and microcanonical variants (Peixoto, 2014, 2019; Côme et al., 2021). By linking community detection with core-periphery decomposition, these models show that communities themselves can act as cores or peripheries relative to one another.

A structure closely related to the multi-core core-periphery organisation is the rich-club or elite-circle structure (Alba and Moore, 1978; Zhou and Mondragón, 2004; Colizza et al., 2006), which describes the tendency of high-degree nodes to form densely interconnected groups. In social networks, this phenomenon often signals the emergence of an "oligarchy" of influential actors who dominate communication, in contrast to decentralised structures formed by loosely connected communities. Unlike the simple degree assortativity of the two-block model, the rich-club effect cannot be explained by degree heterogeneity alone and can arise in both assortative and disassortative networks (Newman, 2002). This distinction links rich-club theory with multi-core structures, as both point to overlapping or interacting dense subgroups within a larger network.

Despite this rich literature, the interplay between community structure and core-periphery organisation has received limited systematic treatment. While some studies have hinted at their coexistence, to the best of our knowledge, no existing framework explicitly identifies core and peripheral communities. To address this gap,

we introduce a novel approach that generalises the traditional two-block, node-level core-periphery model to the community level, enabling the detection of core and peripheral organisation among communities.

Chapter 5

Detecting community-level core-periphery structures

5.1 Motivating example

Our approach is motivated by a case study on co-authorship within the scientific community of Italian academic scholars, as recorded in the Italian Ministry of University (MUR) roster in December 2022. The data concern collaborations in a co-authorship network involving researchers in statistics, sociology and business over 10 years, from 2012 to 2022. A more detailed description of the data collection phases is reported in (De Stefano et al., 2023a; Fabbrucci Barbagli et al., 2025).

The edges in the co-authorship network represent relationships derived from the shared collaborations in the paper-author bipartite network. Hence, an edge between two authors is formed if both are involved in at least one common scientific paper. The edges can carry weights that reflect the number of shared papers, and edge weights indicate the number of shared publications. In this setting, the weight distribution is highly skewed, with a few author pairs accounting for a disproportionate number of collaborations. Moreover, the weights are largely influenced by factors such as project size and journal authorship rules, which may capture institutional or disciplinary proximity rather than relational strength. For these reasons, we base our analysis on the unweighted version of the network.

Although the academic statistics community collaborates across multiple disciplines, we focus on relationships between statistics, sociology, and business to capture the strongest collaborations in applied quantitative research within the social and economic domains. This disciplinary focus is enabled by the availability of detailed information on academic subfields through the Italian *Settore Scientifico Disciplinare* (SSD, i.e. Scientific Discipline Sector) classification system. As shown in previous research (Fabbrucci Barbagli et al., 2025; De Stefano et al., 2023b; De Stefano et al., 2013), Economic Statistics (S/03) and Social Statistics (S/05) are particularly integrated with economics, finance, and social sciences. Moreover, in many cases, statistics is institutionally located within economics or social sciences departments, reinforcing the relevance of this focus. This proximity may enhance scientific collaboration on cross-disciplinary topics and lead to the formation of research groups, even if they

can be characterised by a pronounced hierarchy, as shown in previous research on similar academic communities (De Stefano et al., 2013)

Additional information for our analysis includes author and publication data sourced from Scopus and MUR. In particular, we take into account authors' location (as determined by their affiliation), academic role, and the thematic focus of their publications. We are interested in assessing whether geographical and cultural proximity, shared research interests, as well as academic seniority drive collaboration (Katz, 1994).

Table 5.1 presents the main characteristics of the co-authorship network. The dataset comprises 2,691 academic scholars collaborating on 1,809 papers, with a network density of 0.09%, reflecting sparse collaboration across the network. Statisticians represent 30.7% of the network, and exhibit the highest connectivity and bridging role, with an average degree and betweenness of 5.06 and 5883.28, respectively, and a density of 0.20%, suggesting strong group interconnections. In contrast, researchers from business and sociology are more weakly connected, with edge densities of 0.05% and 0.04%, respectively.

As expected, collaboration patterns vary by academic seniority. Full Professors show the highest centrality, playing a central role in facilitating connections. Collaboration levels appear relatively consistent across regions, with minor variations. Researchers from the Centre and North-East regions report the highest average betweenness, while those from the Islands show the lowest connectivity.

With a transitivity value of 0.26, the network demonstrates a moderate tendency toward community formation. At the same time, the network exhibits a heterogeneous

		N	%	Average Degree	Average Betweenness	Density (%)
	Total	2691		2.34	2917.98	0.09
Field	Statistics	827	30.73	5.06	5883.28	0.20
	Business	821	30.51	1.26	1489.36	0.05
	Sociology	1043	38.76	1.02	1691.31	0.04
Role	Full Professor	748	27.80	3.20	4662.07	0.12
	Associate	1134	42.14	2.25	2696.91	0.08
	Researcher	809	30.06	1.67	1615.27	0.06
Location	North-West	687	25.53	2.23	2546.70	0.08
	North-East	591	21.96	2.29	3251.66	0.09
	Center	689	25.60	2.39	3369.73	0.09
	South	517	19.21	2.49	2618.23	0.09
	Islands	207	7.69	2.28	2442.49	0.09

Table 5.1: Co-authorship network characteristics.

degree distribution, with highly connected nodes that may belong to different communities and create hierarchical patterns. These patterns suggest the coexistence of different meso-scale structures: strong intra-disciplinary ties define cohesive subnetworks, whereas bridging roles of senior academics, statisticians, and certain regions suggest a core-periphery dynamic. Overall, thematic, institutional, and geographical features combine to form a layered architecture in which communities, cores, and hubs jointly shape collaboration and knowledge flow.

Since co-authorship networks inherently reflect the interdependence of groups of collaborators, we ask whether these communities are equally integrated into the broader network or whether some remain relatively isolated, with members collaborating primarily within their own group. If such isolation exists, our goal is to characterise these communities and investigate whether their decentralisation is shaped by factors such as members' institutional location, academic role, or research focus. Our main contribution lies in formalising the problem and showing that it can be addressed as a core-periphery partition of the identified communities. In addition, we develop a framework capable of detecting this community-level core-periphery organisation.

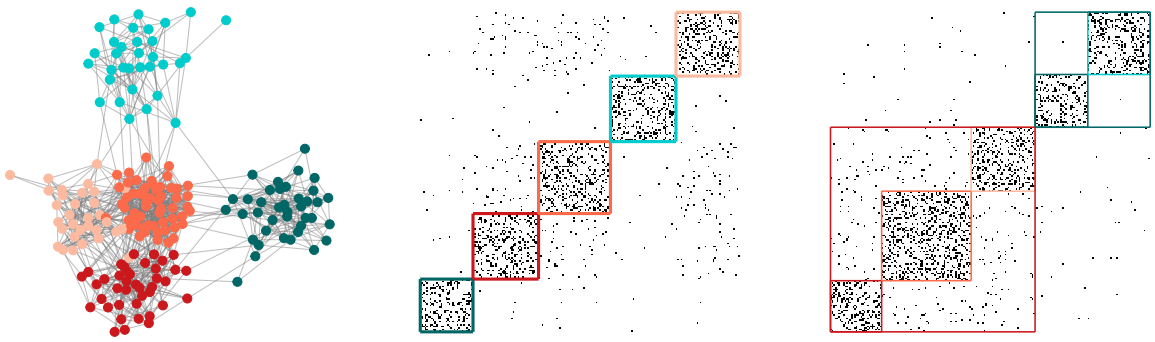
5.2 Methodology

5.2.1 Community-level core-periphery structures

While most analyses focus on the roles of individual nodes, our approach shifts attention to how entire communities of interdependent actors are positioned within broader network architectures. This perspective highlights two nested levels of network structure:

- **Community partition**: Nodes are organised into densely connected groups (communities) that exhibit relatively sparse connectivity between them (Figure 5.1b).
- **Community-level core-periphery partition**: Communities are hierarchically organised, distinguishing those that form the core, which are densely interconnected and central, from those in the periphery, which are sparsely connected and with limited links to other peripheral groups (Figure 5.1c).

The proposed community-level core-periphery classification approach aims to identify groups of nodes based on their relationships and the hierarchical structure among those groups. In many complex social systems, communities can be understood as circles of interdependent actors, and recognising these interdependencies is crucial for: explaining how information, resources, or influence circulate; how collective dynamics are sustained; and how certain groups come to occupy more central positions within the broader network. In this framework, central positions are typically held by communities that are densely interconnected



(A) (B) (C)

Figure 5.1: (a) Example of a network partition in 5 communities with core-periphery roles; (b) Adjacency matrix representing a community partition of the nodes; (c) Adjacency matrix representing a community-level core-periphery partition. Blue colour denotes the periphery, while red colour denotes the core.

and act as bridges across different parts of the system (i.e., core communities). Peripheral communities, by contrast, are weakly connected among themselves, with their participation in the system often mediated by their relations with the core.

Shifting the attention from individual nodes to groups of nodes enables the identification of collective behaviour and shared structural roles. This perspective reflects the fact that many outcomes–such as innovation, diffusion, or resilience–are driven by interactions among groups rather than isolated actors. By treating communities as the analytical unit, we capture meso-scale dynamics that more accurately represent how functional subsystems operate within a network.

The proposed community-level core-periphery classification approach substantially enhances interpretability: by grouping nodes based on shared structural or attribute-based characteristics, it becomes possible to directly relate network positions to real-world social or organisational entities. For example, categorical attributes, such as geographical location or functional role, can naturally serve as cluster labels, allowing us to examine how different types of groups align with core or peripheral positions. This helps identify which subsystems or domains serve as key hubs within the overall structure.

Moreover, by analysing inter-community connection patterns, we can assign and interpret intra-community leadership. This means that the density or strength of inter-community connections helps identify which communities, and which nodes within those communities, act as central influencers.

An additional motivation for identifying core communities stems from empirical observations in SBMs, which, in networks with substantial degree heterogeneity, often reveal clusters with hub-like behaviour, that is, groups that maintain dense connections across the network (Karrer and Newman, 2011). Rather than interpreting these highly connected clusters in isolation, grouping them into a unified core block provides a clearer understanding of their structural role.

Finally, adopting a community-level perspective improves both scalability and robustness in large, complex networks. Aggregating nodes into communities reduces noise from individual-level variability–such as the disproportionate influence of highly connected hubs–and facilitates the detection of persistent structural patterns across scales. This is particularly valuable in networks that span multiple domains, organisations, or regions, where individual-level connectivity may obscure broader systemic dynamics.

5.2.2 Community-level core-periphery classification

In this section, we present the proposed community-level core-periphery classification framework. The framework is based on an objective function specifically designed to identify community-level core-periphery structures under the constraint that the set of communities is partitioned into exactly one core group and exactly one peripheral group, each containing at least two communities. This formulation is inspired by the objective function introduced in Cucuringu et al. (2016), but extends it by replacing node-level coreness with measures that summarise connectivity within and between communities.

The procedure consists of three main steps.

Step 1. Partition the node set of the network into K communities. These communities may be the result of the application of a community detection method appropriate for the data and the specific research task. Alternatively, the communities can correspond to an existing classification of the nodes, for instance, one corresponding to a particular node-level categorical attribute.

Let $Q = \{q_1, \ldots, q_k, \ldots q_K\}$ denote the set of communities.

Step 2. Estimate the connectivity between the each pair of communities q_k and q_l:

$$\hat{\theta}_{kl} = \frac{m_{kl}}{n_k n_l}, \quad k, l = 1, \ldots, K, \quad k \neq l,$$

where m_{kl} is the number of observed edges between the two communities, and n_k, n_l are their sizes (i.e. number of nodes assigned to each community). Note that this estimator is consistent with the connectivity parameter estimates of a standard block model or SBM for binary networks (White et al., 1976; Holland et al., 1983; Karrer and Newman, 2011).

Step 3. Evaluate the quality of candidate core-periphery partitions of the communities. This is achieved by computing the value of an objective function, $\phi(\mathbf{z}; \hat{\boldsymbol{\Theta}})$, where $\hat{\boldsymbol{\Theta}}$ denotes the estimated connectivity matrix with off-diagonal entries $\hat{\theta}_{kl}$ and null diagonal, and $\mathbf{z}$ is an indicator vector specifying whether a each community belongs to the core or the periphery. The function $\phi(\mathbf{z}; \hat{\boldsymbol{\Theta}})$ is then optimised with respect to $\mathbf{z}$ over all admissible partitions of the communities into core and periphery groups.

We now describe Step 3 in more detail. Formally, each community q_k is associated with a binary variable $z_k \in \{0, 1\}$, where $z_k = 1$ if community q_k belongs to the core, while $z_k = 0$ if it belongs to the periphery.

We impose the constraints:

$$\sum_{k=1}^{K} z_k \geq 2 \quad \text{and} \quad \sum_{k=1}^{K} (1 - z_k) \geq 2,$$

ensuring that both the core and the periphery contain at least two communities. This excludes degenerate cases with core or periphery comprising only one community. The partitioning of Q into core and peripheral communities is thus determined by the vector $\mathbf{z} = (z_1, \ldots, z_k, \ldots, z_K)$.

The connectivity strengths between pairs of communities, captured in $\hat{\boldsymbol{\Theta}} = [\hat{\theta}_{kl}]$, reflect both the topology (presence or absence of edges) and the intensity (edge frequency) of inter-community interactions. To evaluate a given partition $\mathbf{z}$, we define the following sets of nonzero inter-community connectivity values extracted from $\hat{\boldsymbol{\Theta}}$, which encode connectivity among core and peripheral communities, respectively:

$$\mathcal{T}_c = \left\{ \hat{\theta}_{kl} : k < l,\ z_k z_l = 1,\ \hat{\theta}_{kl} > 0 \right\},$$

$$\mathcal{T}_p = \left\{ \hat{\theta}_{kl} : k < l,\ (1 - z_k)(1 - z_l) = 1,\ \hat{\theta}_{kl} > 0 \right\}.$$

The objective function $\phi(\mathbf{z}; \hat{\boldsymbol{\Theta}})$ is defined as

$$\phi(\mathbf{z}; \hat{\boldsymbol{\Theta}}) = \left(\delta_c + \mu_c - \sigma_c\right) - \left(\delta_p + \mu_p + \sigma_p\right), \tag{5.1}$$

where:

- δ_c, δ_p denote the densities of nonzero edges in the core and periphery, respectively:

$$\delta_c = \frac{|\mathcal{T}_c|}{\sum_{k<l} z_k z_l}, \qquad \delta_p = \frac{|\mathcal{T}_p|}{\sum_{k<l} (1 - z_k)(1 - z_l)};$$

- μ_c, μ_p are mean edge frequencies in the core and periphery,

$$\mu_c = \frac{1}{|\mathcal{T}_c|} \sum_{\hat{\theta}_{kl} \in \mathcal{T}_c} \hat{\theta}_{kl}, \qquad \mu_p = \frac{1}{|\mathcal{T}_p|} \sum_{\hat{\theta}_{kl} \in \mathcal{T}_p} \hat{\theta}_{kl};$$

- σ_c, σ_p are standard deviations (sd) of the edge frequencies,

$$\sigma_c = \sqrt{\frac{1}{|\mathcal{T}_c| - 1} \sum_{\hat{\theta}_{kl} \in \mathcal{T}_c} (\hat{\theta}_{kl} - \mu_c)}, \qquad \sigma_p = \sqrt{\frac{1}{|\mathcal{T}_p| - 1} \sum_{\hat{\theta}_{kl} \in \mathcal{T}_p} (\hat{\theta}_{kl} - \mu_p)}.$$

Zero entries are excluded from $\mathcal{T}_c$ and $\mathcal{T}_p$ to ensure that the summary statistics capture only the strength and variability of existing inter-community connections, while the overall proportion of such connections is already reflected in the density terms. This choice is in line with the definition of periphery, where links among peripheral communities are expected to be sparse, and when present, as weak as possible. By considering only nonzero entries, the method is able to identify configurations in which peripheral links exist but remain weak, thereby emphasising interaction intensity rather than sole presence or absence.

By construction, the function in Equation (5.1) increases when core communities are denser, more strongly connected, and more homogeneous than peripheral ones. A high value of δ_c and a low value of δ_p indicate a denser core and a sparser periphery, consistent with a typical core-periphery structure. We use the mean (μ) as a measure of the central tendency of the edge frequency distribution. Maximising μ_c while minimising μ_p reinforces the contrast between the interaction intensity within the core and that within the periphery. We use the standard deviation (σ) as a measure of dispersion. The σ_c and σ_p terms penalise high variability in edge frequencies, preventing solutions in which the core consists only of a few unusually strong links and the periphery of links of very heterogeneous strength. This also discourages the opposite case, with a small periphery with extremely weak links and a core with heterogeneous interconnections.

The groups are determined endogenously by the data, with no constraints on their size. We do not require balanced groupings, and the number of core and peripheral communities can vary as long as both sets include at least two elements.

The optimisation of the function in Equation (5.1) with respect to $\mathbf{z}$ is a combinatorial problem, as it requires evaluating discrete partitions of K communities into core and periphery. The search space of admissible binary partitions grows exponentially as $O(2^K)$ in the worst case, making exhaustive enumeration infeasible even for moderately sized networks. To address these challenges, we use a genetic algorithm (Scrucca, 2013), inspired by natural selection. Genetic algorithms are particularly well-suited to this problem because they can efficiently explore large and complex search spaces, balance exploitation of promising solutions with exploration of new ones, and avoid getting trapped in suboptimal local optima. This approach allows us to approximate the optimal community-level assignment between the core and the periphery, while maintaining computational feasibility and robustness across different network structures.

To assess the empirical performance of the proposed approach, we conduct a simulation study based on varying networks generated from SBMs. As highlighted in the introduction, while the SBM is widely recognised for its ability to generate networks with community structures, it is also sufficiently flexible to accommodate more complex patterns, including core-periphery configurations. In particular, the generative model developed for this study is explicitly designed to simulate networks where both community structure and inter-community heterogeneity coexist.

We aim to evaluate how effectively the proposed approach can recover the true

partitioning of communities into core and periphery, both in terms of objective function behaviour and core-periphery classification performance.

5.3 Simulation study

5.3.1 Simulation design

Networks are generated to reproduce realistic patterns of inter-community collaboration. Different network sizes, $n = \{100, 200, 500, 1000\}$ and numbers of communities, $K = \{5, 10, 20\}$ are considered. A parameter $\lambda = \{25, 50, 75\}$ controls the percentage of communities belonging to the core. The remaining communities are treated as peripheral. Each node is then assigned to one of these communities according to a probability distribution, which is either uniform (equally sized communities) or drawn from a Dirichlet distribution to reflect size heterogeneity across communities.

Once both the community and community-level core-periphery membership are determined, we define connection probabilities within and between communities. Nodes belonging to the same community are connected with a high probability. Inter-community connections are formed based on the core-periphery classification of the involved communities: core-to-core connections are most likely, core-to-periphery connections are less likely, and periphery-to-periphery connections are least likely. Further details are provided in the Appendix E.1.

To ensure realistic scenarios, we exclude certain parameter combinations that could produce degenerate structures: specifically, small networks (with 100 or 200 nodes) were not paired with a large number of communities ($K = 20$), and large networks (with 500 or 1000 nodes) were not paired with very few communities ($K = 5$). The collection of parameter combinations provides 96 distinct scenarios; for each scenario, we simulate 100 independent network data realisations. Data are generated, ensuring that every community includes at least two nodes.

The performance of core-periphery classification at the node level is evaluated using standard metrics: balanced accuracy (BA) results are presented in the main text, while F1 scores are provided in Appendix E.3. Additional results on the estimation of the number of node-level clusters are also provided in Appendix E.3.

5.3.2 Simulation results: objective function behavior

Figure 5.2 shows the results of the evaluation of the behaviour of the proposed objective function. The analysis is conducted on synthetic networks with $n = 1000$ nodes, $K = 20$ equally sized communities, and known core-periphery structures $\mathbf{z}$ under three different λ values. Objective function values $\phi(\mathbf{z}; \boldsymbol{\Theta})$ are computed for all possible configurations of $\mathbf{z}$, using the actual data-generating $\boldsymbol{\Theta}$ values.

The BA measures the agreement between detected and actual $\mathbf{z}$. Blue points represent solutions with more peripheral communities than the true structure, while

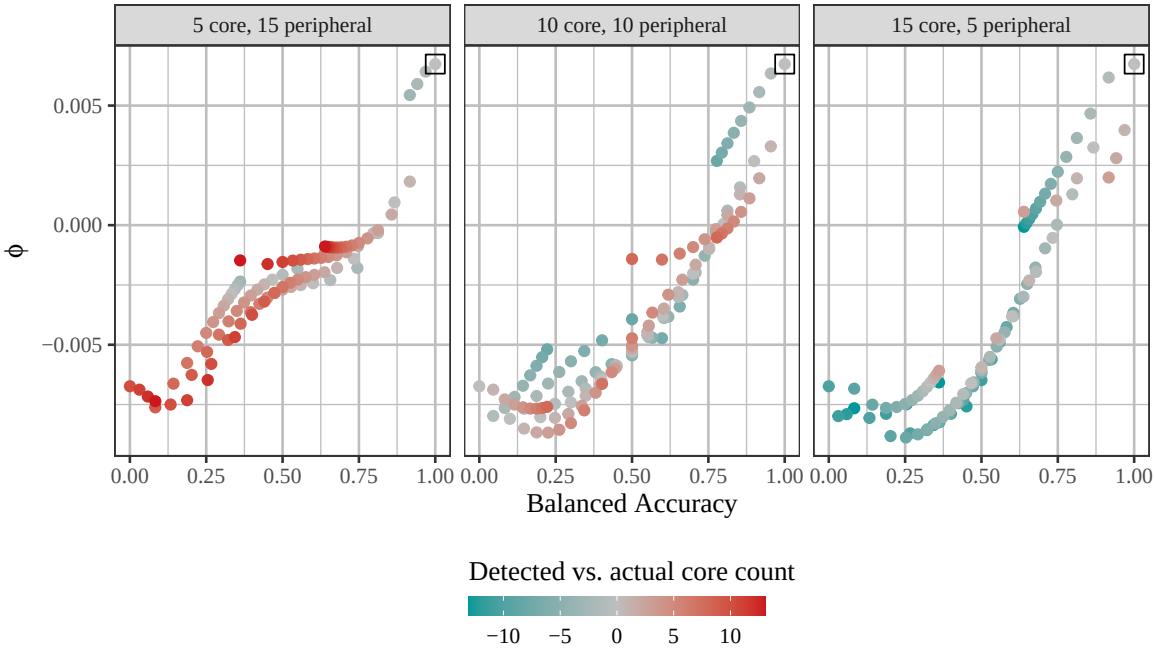


Figure 5.2: Scatter plot of objective function (ϕ) vs. balanced accuracy (BA) scores for core-periphery networks with $n = 1000$ nodes, $K = 20$ communities, and core proportions of 25%, 50%, and 75% (panel from left to right). Blue points represent solutions with more peripheral communities than the true structure, while red points indicate solutions with more core communities. The true solution (BA = 1) is highlighted with a square.

red points indicate solutions with more core communities. The true configuration (BA = 1) is highlighted with a square. The objective function ϕ reaches its maximum at the true solution, while close but suboptimal values are observed when one additional peripheral community is included in each of the three different configurations.

5.3.3 Simulation results: core-periphery classification performance

The second simulation study investigates how structural parameters, such as network size, number (K), and size of communities, and the proportion of core groups, influence the ability to detect core-periphery structures at the community level.

As shown in Figure 5.3 and 5.4, our method performs robustly across a wide range of realistic network configurations. Since the approach assumes a fixed community partition as input, its performance depends on the quality of this initial grouping. To assess this, we apply the method following three standard node-level community detection algorithms: the binary SBM, implemented via the *blockmodels* package (Leger et al., 2021b), and the Louvain and Infomap algorithms, both available in the igraph package (Csárdi et al., 2025; Csárdi and Nepusz, 2006).

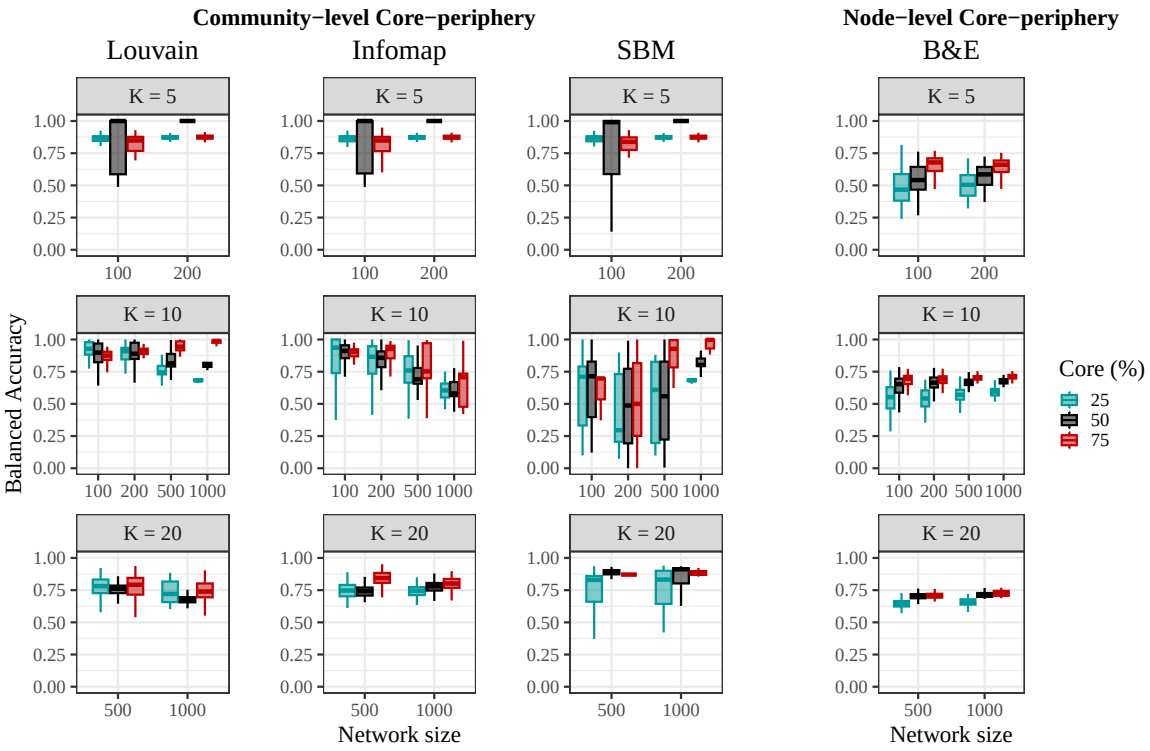


*FIGURE 5.3: **Uniform** community sizes – Balanced accuracy distribution across different network sizes and numbers of communities (K). Results are shown for different community detection methods, including Louvain, Infomap, and SBM, as well as for the Borgatti and Everett (B&E) approach.*

Overall, the proposed method demonstrates superior balanced accuracy for communities identified by the binary SBM compared with alternative approaches, particularly in networks with moderate to large sizes ($n = 500$ or 1000). This is especially pronounced when the proportion of core communities is high (75%), and the community sizes are non-uniform, reinforcing its suitability for complex and realistic network structures. When node-level community detection is implemented using Louvain, performance tends to be higher in smaller networks and under conditions of equal-sized communities. When using Infomap for community detection, which emphasises flow-based modular structures, the proposed approach tends to yield the lowest overall performance in large networks and in non-uniform scenarios.

Across all algorithms, the proportion of core communities influences performance. In small networks, optimal performance is achieved at an intermediate core proportion (50%), whereas in larger networks, performance peaks at 75%. However, it is important to note that in scenarios with five communities, a core proportion of 25% or 75% results in only one core or one peripheral community, respectively. These are degenerate cases that pose a challenge for our method, which by design requires at least two communities in both the core and the periphery to define a valid core-periphery

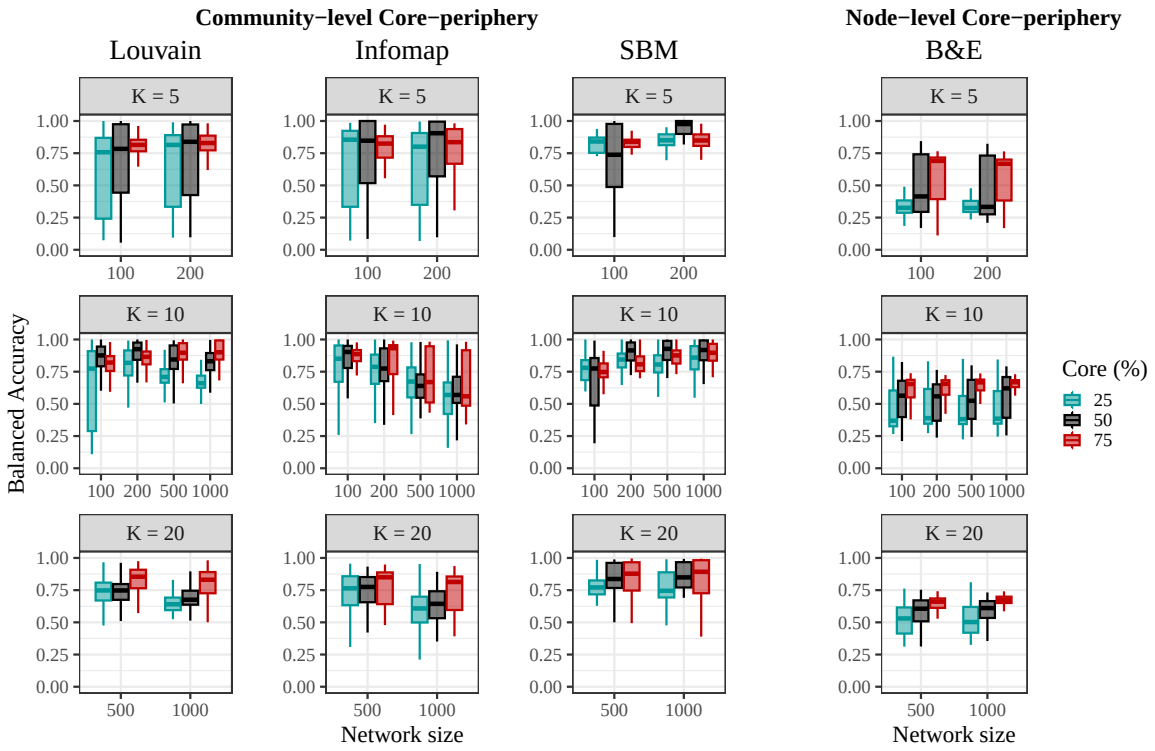


*FIGURE 5.4: **Non-uniform** (Dirichlet-distributed) community sizes – Balanced accuracy distribution across different network sizes and numbers of communities (K). Results are shown for different community detection methods, including Louvain, Infomap, and SBM, as well as for the Borgatti and Everett (B&E) approach.*

partition. Consequently, the method is not capable of identifying the true solutions in these settings, which explains the observed lower performance. Nevertheless, in small networks with equally sized communities, the method shows strong potential when used in combination with all three community detection algorithms.

For benchmarking purposes, we also compare our community-level approach with the node-level core-periphery detection model proposed by Borgatti and Everett, which remains a standard reference in this area. The method was implemented using the netUtils package in R (Schoch, 2024). As shown in Figure 5.3 and 5.4, the Borgatti-Everett model consistently achieves lower BA across all scenarios. This difference is particularly pronounced in networks with a low proportion of core communities, where the node-level approach fails to capture higher-order structural regularities. These results highlight the importance of leveraging community partitions to reliably identify community-level core-periphery organisation at the meso-scale.

5.4 Application: Italian academics co-authorship network

The case study on co-authorship among Italian academic scholars illustrates the practical application of the proposed methodology. We examine two complementary settings. In the first, we infer node-level communities defined by high internal connectivity, with the objective of uncovering collaboration patterns at the researcher level. In the second, we assign community labels based on the regional affiliation of authors, allowing us to investigate collaboration dynamics and core-periphery structures across Italian regions. The underlying hypothesis is that geographical proximity may foster core regional clusters and that the emergence of core and peripheral regions may further reflect thematic or institutional specialisation within the academic system.

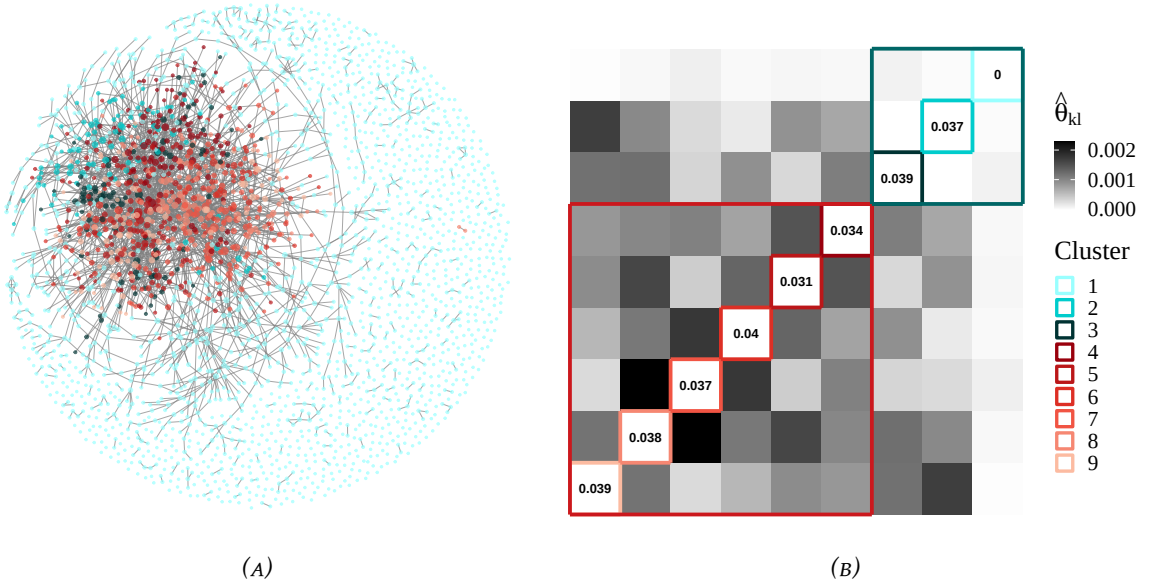
5.4.1 Communities detected using SBM

We begin by applying the degree-corrected SBM (Karrer and Newman, 2011) to the co-authorship network in order to obtain a meaningful community partition. This community detection approach is particularly suitable for this context, as it accounts for the heterogeneous degree distribution observed in the network, while still detecting assortative block structure.

Importantly, the model is fitted to the entire network, explicitly allowing for isolated nodes (i.e., nodes with degree zero) and disconnected components. To assess the robustness of the results, we also applied the same procedure to the giant component only. The resulting partition shows almost perfect agreement with that obtained on the full network. For completeness, we therefore report the results based on the entire network.

The degree-corrected SBM identifies eight cohesive communities and one large group of sparsely connected or isolated nodes, which contains 70.6% of all network nodes (light blue in Figure 5.5). Within this group, 66.9% of nodes are disconnected from the largest component of the network, while only two interconnected but otherwise isolated nodes fall outside it. The distribution of the sizes of the remaining eight clusters is relatively uniform, ranging from 86 to 116 nodes.

From this community structure, we compute the estimated connectivity matrix $\hat{\boldsymbol{\Theta}}$ as described in Section 5.2.2, with entries capturing the edge density between communities. Using this matrix, our method evaluates all admissible partitions of the communities into core and periphery by optimising the objective function $\phi(\mathbf{z}; \hat{\boldsymbol{\Theta}})$. This procedure classifies the detected communities into six core and three peripheral groups. The resulting core set corresponds to densely connected and structurally homogeneous communities, whereas the peripheral set consists of groups characterised by sparse and heterogeneous connectivity patterns. Notably, the latter also includes the large light blue community, which contains the network's isolated nodes.



(A) (B)

FIGURE 5.5: (a) Co-authorship network partition in communities and their core-periphery organisation. (b) Co-authorship connectivity matrix ($\hat{\Theta}$) representing the community-level core-periphery partition. Blue square denotes intra-periphery connectivity, red square intra-core connectivity.

Results are presented in Figure 5.5. The left panel shows the community partition of the network, and the right panel displays the connectivity matrix $\hat{\Theta}$ representing the community-level core-periphery partition. Each diagonal element of this matrix represents a community.

To better reveal inter-regional links, we visualise the off-diagonal values in the $\hat{\Theta}$ matrix, as a few highly assortative clusters produce dark diagonal cells that obscure off-diagonal values in the full heatmap. Interaction probabilities between peripheral communities are nearly zero, as indicated by the white cells in the blue square (intra-periphery block), and their links to core communities are weak. In contrast, core communities exhibit more dense, though still relatively weak, interactions, represented by the light grey cells in the red square (intra-core block).

For comparison, we also applied the node-level core-periphery detection model of Borgatti and Everett (2000). This approach results in a trivial partition of the network into high- and low-degree nodes, essentially reproducing degree heterogeneity: low-degree nodes naturally appear as peripheral (Table 5.2). To assess the consistency between our community-level core-periphery assignment and the node-level B&E partition, we compared the two partitions on the co-authorship network. The results show a BA of 0.78 and an F1 of 0.89, indicating partial agreement between the two approaches but also meaningful differences in how the "core" role is defined. Most nodes identified as peripheral by our method (2,016 out of 2,483) are also classified as peripheral by the B&E model, while 67 are labelled as core. Conversely, among

the nodes belonging to core communities in our framework, 190 (31%) are also assigned to the B&E core, whereas 418 (69%) are classified as peripheral. This pattern suggests that our method identifies additional locally central actors–those having dense connections within a core community–who may not appear as globally central in the B&E sense, consistent with the community-level focus of our approach.

The distinction between core and periphery structure may be explained by thematic specialisation. To characterise the thematic content of each community and of the core and periphery, we combine term frequency (tf) and term frequency-inverse document frequency (tf-idf) statistics Robinson (2017), computed from words appearing in titles and keywords, with data on the frequency of publication venue (scientific journals). These quantities were computed in R using the functions implemented in the tidytext package (Robinson and Silge, 2025). The tf results are presented in Figure 5.6 in the main text, while tf-idf results can be found in Figure E4 of Appendix E.4.

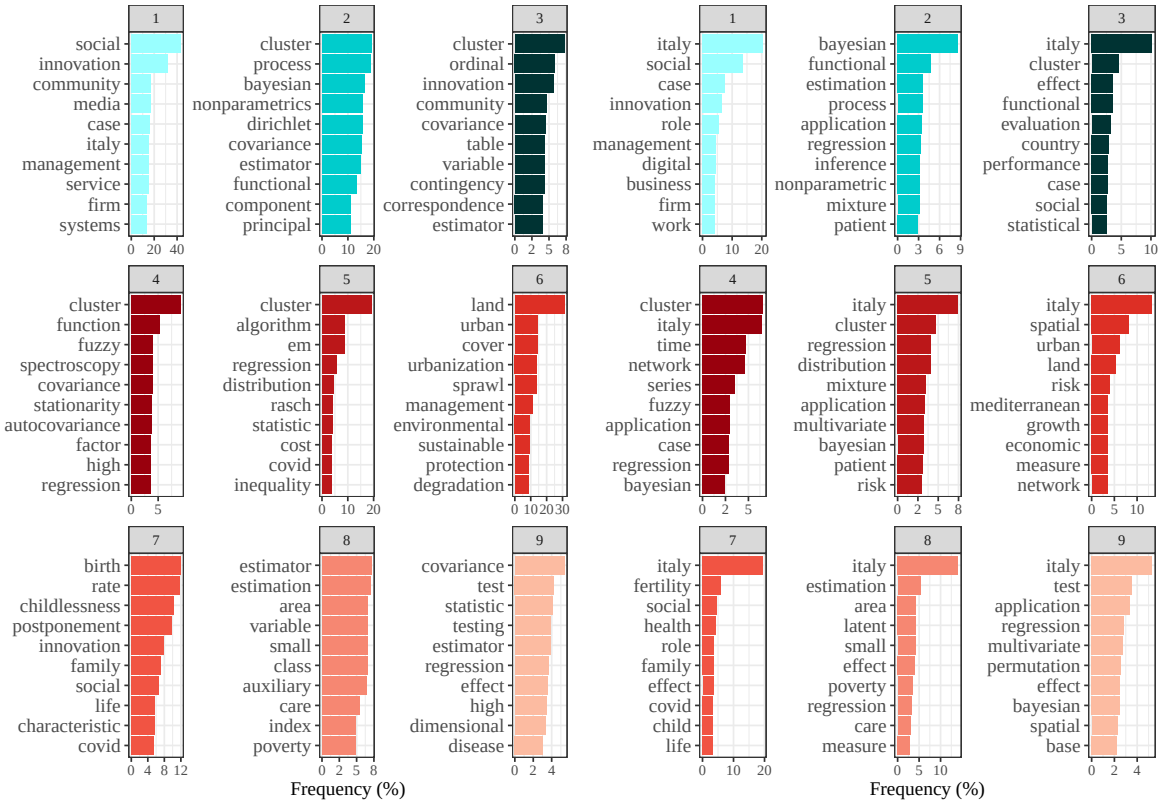
Standard text preprocessing steps were applied to ensure that only representative words are retained. Specifically, we split the texts into individual words (word tokenisation) and lemmatised them to reduce them to their unique dictionary form. Next, we removed stop-words (i.e., prevalent and non-informative words such as "and", "the", and "to"), numbers, and symbols. The data were further cleaned to exclude extremely common but non-informative words, such as "data", "model", "analysis", and "approach".

The largest peripheral cluster (Cluster 1) is dominated by themes from the social sciences and applied fields, as reflected in frequent terms such as "social", "innovation", and "management", and in high tf-idf terms such as "democracy", "populism", and "luxury". It is also associated with various Italian sociological journals, including *Italian Sociological Review*, *Partecipazione e Conflitto*, and *Italian Journal of Sociology of Education*. Furthermore, a substantial proportion of the publications in this cluster appeared in more recent years, suggesting a relatively young and still consolidating research area, characterised by limited collaborations.

The second peripheral cluster (Cluster 2) centres on a highly specialised domain of statistical theory and methodology. Distinctive terms ("bayesian", "nonparametrics", "dirichlet", "pitman", "gibbs") and high-impact journals (i.e. *Biometrika*, *Electronic Journal of Statistics*, and *Journal of the American Statistical Association*) indicate that this cluster represents advanced knowledge production in complex statistical methodology. Its theoretical orientation and narrow focus further characterise it as peripheral in the overall collaboration structure.

Table 5.2: Degree distribution by core membership assigned using Borgatti and Everett model.

	n	Mean	SD	Median	Min	Max
Periphery	2434	1.53	1.66	1	0	6
Core	257	9.96	3.66	9	7	30



(A) Paper topic *(B) Paper title*

FIGURE 5.6: (a) Distribution of topic term frequency (tf) by cluster and core-periphery organisation. (b) Distribution of title tf by cluster and core-periphery organisation.

The third peripheral group (Cluster 3) integrates applied statistics with a wide range of application areas. Dominant tf terms include "ordinal", "effect", "covariance", and "country", while tf-idf highlights terms such as "symbolic", "interlock", "histogram", and "port". Journals like *Electronic Journal of Applied Statistical Analysis*, *Quality and Quantity*, as well as field-specific outlets such as *Social Indicators Research*, *British Food Journal*, and *Public Health Nutrition*, position this cluster within applied quantitative methods for social and economic research, with a notable subfield focusing on nutrition and food related health studies. The peripheral role of the cluster reflects its predominantly applied nature and cross-disciplinary involvement. Rather than forming integrated collaborative groups, they contribute analytical expertise in different thematic areas, giving rise to weaker and more transient collaborations.

Overall, the three peripheral clusters comprise scholars from distinct scientific domains, each focused on research topics that frequently require collaboration with researchers possessing complementary–rather than thematically aligned–expertise. Based on the journal distribution reported in Figure 5.7, Cluster 1 is predominantly composed of sociologists and economists working on themes such as sustainability,

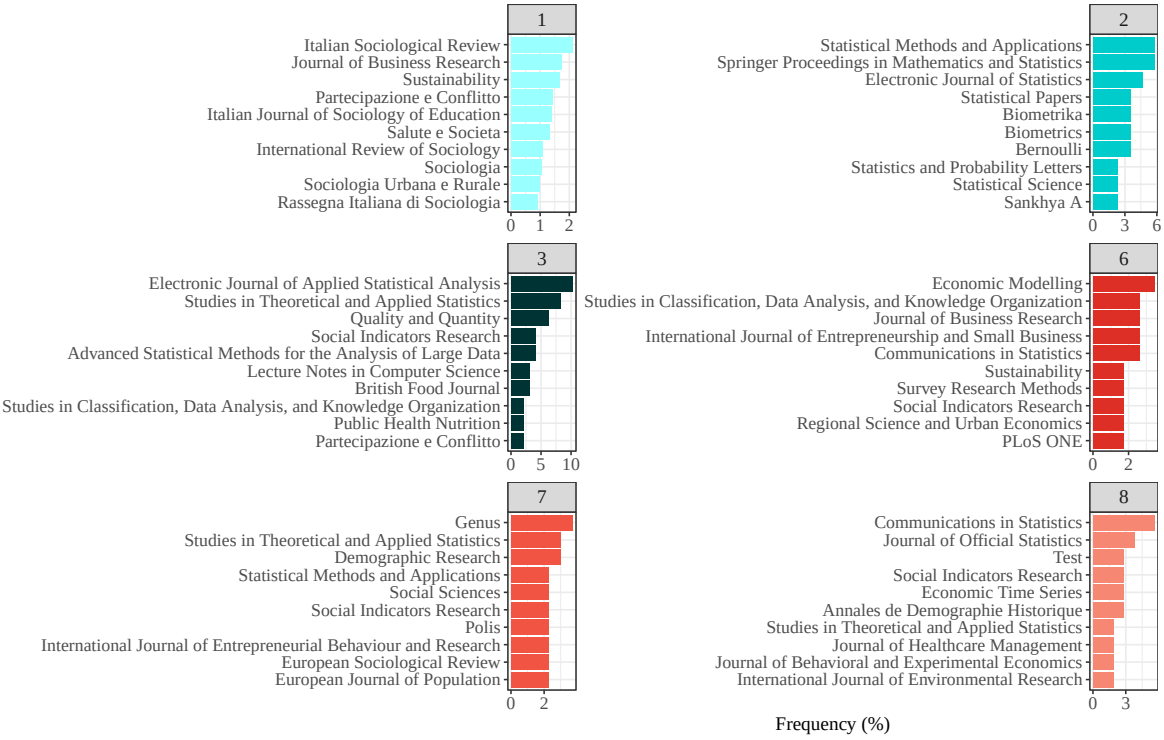


FIGURE 5.7: Distribution of journal frequency by cluster and core-periphery organisation.

political studies, urban and regional sociology, and education, and exhibits a generally low propensity to collaborate. Clusters 2 and 3 comprise two distinct profiles of statisticians whose reciprocal collaboration networks are comparatively limited: the former due to their focus on applied research within specialised domains, and the latter because of their strong orientation toward theoretical topics. As a result, they tend to collaborate mainly with statisticians located in the core communities. For example, members of Cluster 2 often work with environmental and medical researchers (mainly located in the core Cluster 9 as described below) rather than extensively within their own clusters. These peripheral communities are the only groups in the network that display almost no interaction with one another.

In contrast, the six core clusters are characterised by greater thematic homogeneity. Cluster 4 centres on methodological innovation in time-series analysis, and financial and economic modelling, featuring terms such as "time series", "autocovariance", "stationarity"). Cluster 5 emphasises general statistical modelling and inference, with distinctive terms including "em algorithm", "mixture", "symmetric", "invariance", "irt". Clusters 6 and 7 are strongly disciplinary. Cluster 6, linked to journals such as *Economic Modelling* and *Regional Science and Urban Economics*, is oriented toward land use and environmental studies (frequent terms include "land", "urban", "spatial", "sprawl", "degradation", "desertification"). In contrast, Cluster 7 is rooted in demography and family studies, featuring distinctive terms such as "birth rate",

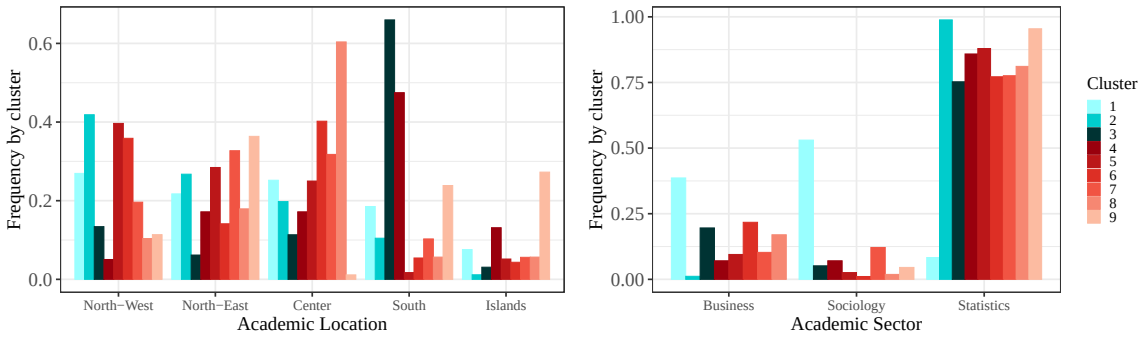


Figure 5.8: (a) Geographic distribution of authors by cluster and core–periphery organisation.(b) Academic sector distribution of authors by cluster and core–periphery organisation.

"fertility", "childlessness", and "grandchild", and characterised by journals including *Demographic Research* and *European Journal of Population*. Cluster 8 partially overlaps thematically with Cluster 7, with terms like "small area", "index", "poverty", and "parturition", and is characterised by journals such as the *Journal of Official Statistics* and *Health Policy*. Cluster 9 combines applied statistics with biomedical and environmental applications (reflected in terms such as "disease", "effect", "asthma", "seismicity") with a strong link to environmental and health research, represented by journals such as *Ecological Indicators*, *Statistics in Medicine*, and *Science of the Total Environment*. The members of the clusters detected as core tend to collaborate repeatedly within well-established research domains, forming dense, stable ties that sustain the network's structural cohesion.

Figure 5.8 presents information on researcher characteristics, including the distribution of authors' geographic locations and authors' academic sectors by cluster. Peripheral cluster 3 shows a marked predominance of researchers based in southern Italy (66.0%), compared with a much smaller southern presence in the core clusters, where it rarely exceeds 20% (Figure 5.8a). Cluster 2 is dominated by academics from the North-East and North-West, clearly associated with the Bayesian non-parametric groups centred at the University of Padova and Bocconi University.

As expected from exploratory analyses, the large peripheral Cluster 1 groups together most of the non-statisticians, who are only weakly connected to the rest of the network (Figure 5.8b). 91.7% of the nodes in that community belong to the business or social sciences, highlighting the bridging role of statistics as a multidisciplinary field. Notably, the overall core-periphery connectivity is guided by statisticians. This high propensity to collaborate is in line with previous studies on the Italian Statistician community (Bacci et al., 2023; De Stefano et al., 2023b).

Overall, the results confirm that peripheral clusters not only diverge thematically but also have distinctive social and institutional compositions. To further characterise and interpret the core and periphery groups, we identified leaders within each

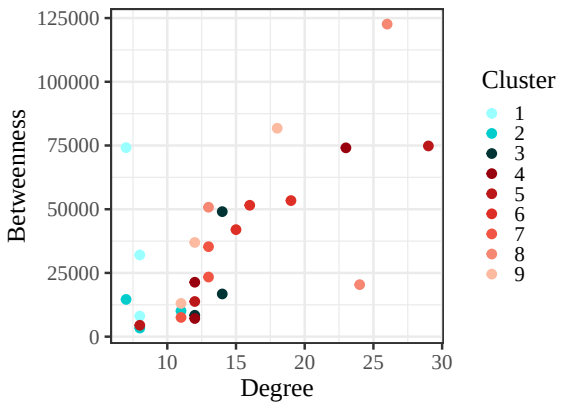


FIGURE 5.9: Top three leaders' degree and betweenness centrality for core and peripheral clusters.

community as the three members with the highest number of inter-community collaborations (a total of 27 leaders, 9 periphery leaders, and 18 core leaders). In the case of ties, degree centrality and then betweenness centrality were used as tie-breakers. Core leaders consistently display higher values across all centrality measures, reflecting their more central and well-connected positions in the collaboration network (Figure 5.9). On average, their degree is 15.9, compared to 9.9 for peripheral leaders, and their mean betweenness centrality reaches 40,800, nearly double that of peripheral leaders (24,100), indicating a greater role in bridging collaborations across communities.

Leaders are predominantly full professors, concentrated in central and northern regions such as Lazio, Lombardia, and Toscana, and are exclusively affiliated with the Statistics sector (Table 5.3). Periphery leaders generally exhibit a more heterogeneous mix of academic ranks (including researchers and associates), and a broader geographic distribution, extending to southern regions such as Campania, Puglia, and Calabria. The only community led by scholars outside the field of statistics (i.e., sociology) is peripheral.

5.4.2 Communities defined by geographical regions

Using the categorisation of the 20 Italian regions as the node-level community partition, the estimated average connection probability is 0.027 within clusters and 0.00039 between clusters, revealing a strongly assortative regional structure. The proposed method identifies 14 core and 6 peripheral regions: Basilicata, Friuli Venezia Giulia, Liguria, Molise, Puglia, and Valle d'Aosta. Highly represented regions (over 250 authors) are all assigned to the core, whereas those scarcely represented (fewer than 15 authors) are in the periphery. One determinant of the core-periphery organisation is the geographical proximity, and our method allows us to detect which regions have few or no connections reciprocally.

Figure 5.10 shows very weak intra-periphery interactions (top-right blue square) and sparser core-periphery links compared to core-core links (bottom-left red square). For each region, the three leaders were identified as outlined above (a total of 59 leaders, 17 representing peripheral regions and 42 core regions, and characterised in Table 5.4). Core regional leaders show higher centrality values, with an average degree of 11.62 compared to 5 for peripheral region leaders and an average betweenness of 26584.79 compared to 4252.81 for peripheral region leaders, indicating their role as connectors in the collaboration network. They are predominantly full and associate professors in the Statistics sector. In contrast, peripheral regional leaders have a greater representation of researchers and associates, and show greater disciplinary diversity, with some leaders affiliated with the business field (e.g., all leaders from Basilicata and Valle d'Aosta).

5.5 Final remarks

We introduced a community-level approach to core-periphery classification, designed to distinguish densely connected core groups from sparsely connected peripheral ones based on both the density and strength of inter-community interactions. By formulating an objective function tailored to this task, our method provides a scalable, interpretable, and flexible framework applicable to a variety of real-world networks. The proposed methodology is showcased to analyse a collaboration network of Italian academic scholars.

		Periphery	Core
Total		9	18
Field	Statistics	7	18
	Business	0	0
	Sociology	2	0
Role	Full Professor	4	11
	Associate	4	5
	Researcher	1	2
Location	North-West	0	5
	North-East	3	1
	Center	1	8
	South	5	3
	Islands	0	1

TABLE 5.3: Frequency of leaders by sector, role, and location across core and peripheral clusters.

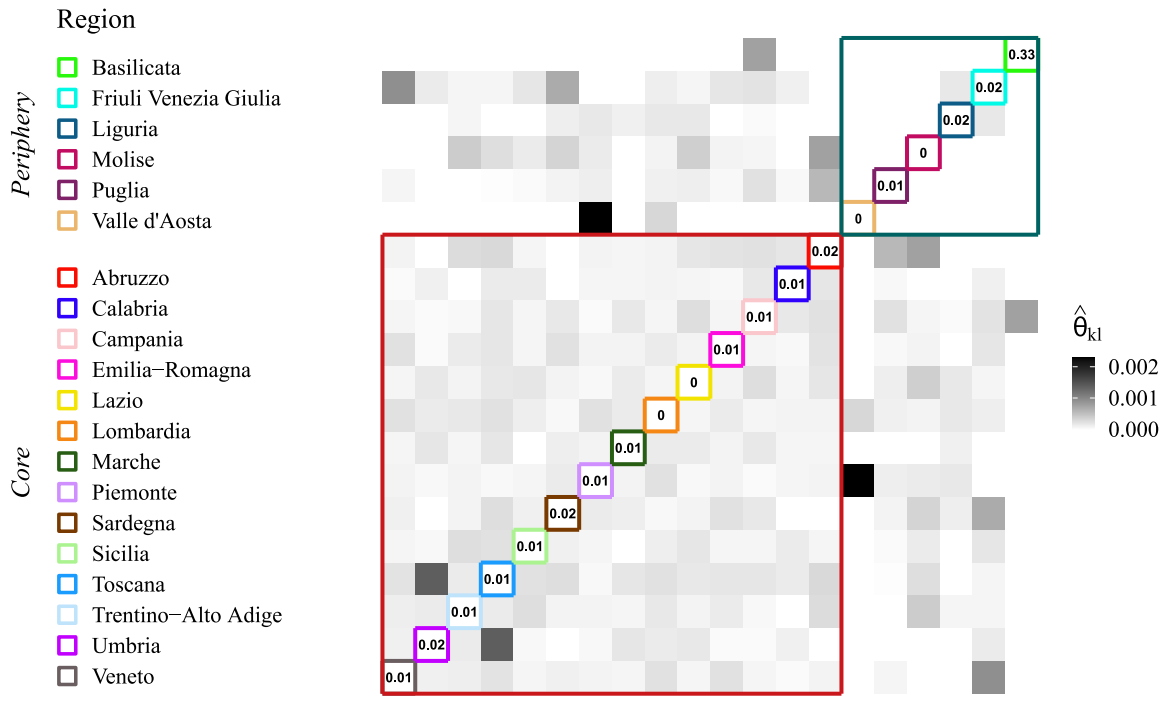


FIGURE 5.10: Regional co-authorship connectivity matrix ($\hat{\Theta}$) illustrating the core-periphery partition of the Italian regions. Blue square denotes intra-periphery connectivity, red square intra-core connectivity.

Our results highlight how the proposed framework departs from classical node-level models (e.g., Borgatti and Everett (2000)) by reconceptualising the periphery. Rather than representing a residual set of loosely connected nodes, the periphery may also include cohesive and internally dense communities that remain weakly integrated into the broader network, tending to form their ties primarily with the

		Periphery	Core
Total		17	42
Field	Statistics	12	40
	Business	5	0
	Sociology	0	2
Role	Full Professor	6	23
	Associate	7	15
	Researcher	4	4

TABLE 5.4: Frequency of regional leaders by sector and role across core and peripheral clusters.

core. This insight helps explain why peripheries can be both intellectually rich and structurally isolated, as shown in our empirical application to the Italian academic collaboration network.

The findings underscore the potential of community-level core-periphery analysis to support science policy. Identifying peripheral groups and their leaders can inform initiatives to strengthen inter-community collaborations, broaden participation across geographic and disciplinary boundaries, and promote a more balanced distribution of resources or enhance scientific innovation, fostering, for instance, preriphery-to-periphery collaboration. More broadly, mapping nested structures enriches our understanding of how influence, knowledge flows, and collaboration dynamics are organised at the group level, complementing established node-level approaches (Zelnio, 2012; Karlovčec et al., 2016; Sedita et al., 2020; Wedell et al., 2022).

From a methodological point of view, an interesting avenue for future extension concerns the choice of summary statistics in the objective function. Our formulation uses mean-based measures, which guarantee a single optimum and stable results. Using more robust alternatives, such as the median, could produce multiple optima, reducing stability but potentially increasing robustness to asymmetric connection densities, where mean-based measures are sensitive to outliers. For a brief discussion and an example, see Appendix E.2.

At the same time, several limitations point toward directions for future work. First, in the current framework, clustering and core-periphery detection are performed sequentially. While this modularity simplifies computation and interpretation, it makes the core-periphery classification dependent on the initial node-level partition. A natural extension is the development of a nested SBM capable of jointly estimating node-level communities and community-level core-periphery roles, thereby integrating the two levels into a unified inferential framework. This direction is inspired by hierarchical SBM approaches (Peixoto, 2019; Côme et al., 2021), though adapted to our specific two-level core-periphery structure. Second, categorical attributes (e.g., role, research interest, location) are used mainly to interpret core-periphery roles and, when treated as community labels, to assess how different group types align with core or peripheral positions. Future work should embed both categorical and continuous attributes directly in the detection step, for example, through attributed network clustering algorithms (Zhang et al., 2023) or attributed block models (Stanley et al., 2019), to enable a richer account of how social and institutional factors shape structural positions. Third, our empirical analysis focuses on a specific small academic community in Italy, constrained by data availability. The method should be tested on larger, more diverse scientific fields and extended to temporal networks (Balland et al., 2019b), to capture how core-periphery roles evolve as collaborations expand and reorganise. Lastly, the application to the co-authorship network considered only binary connections, disregarding edge weights and thereby omitting potentially useful information about collaboration intensity. The proposed framework can, however, be readily extended to weighted networks, particularly in the node-level community detection phase. For example, existing approaches such as weighted SBMs

or modularity-based methods for weighted networks can be directly incorporated into our framework (Fortunato and Hric, 2016; Ng and Murphy, 2021). Extending the core-periphery classification step to weighted settings would require adapting the objective function to account for weighted connectivity patterns (Borgatti and Everett, 2000; Tudisco and Higham, 2019; Rombach et al., 2012). Nevertheless, this remains a more challenging task, as weight distributions in empirical networks are often highly skewed, over- or under-dispersed, and influenced by exogenous factors, such as visibility, team size, or resource availability, rather than purely structural position.

In summary, the present work provides the first explicit framework for detecting nested core-periphery structures at the community level. Future extensions will broaden its scope and deepen its explanatory power, contributing to the broader study of hierarchical organisation in complex networks.

Conclusion

Understanding how groups form in complex networks represents a central question across a wide range of disciplines, from sociology to economics, from urban geography to the study of scientific collaborations, where patterns of interaction among scientific actors have long served as a key empirical setting for investigating community formation, leadership, and hierarchical organisation (Fortunato et al., 2018). In many applied contexts, the goal of network analysis does not merely consist in identifying sets of densely connected nodes, but rather in understanding the mechanisms that drive aggregation, the role played by structurally prominent actors, and the way these dynamics translate into broader forms of hierarchical organisation.

This thesis contributes to the growing literature on community detection and network organisation by proposing an integrated perspective on group formation in attributed networks. Rather than treating leader influence, attribute-based homophily, and center-periphery organization as competing or conceptually distinct mechanisms, the work draws attention to how these operate jointly as interdependent components of a unified structuring process. By explicitly connecting node-level prominence with community-level organisation, the proposed framework extends existing methodological approaches and provides new interpretative tools for understanding how groups emerge and interact in complex networks. In doing so, the thesis not only consolidates insights from previously separate strands of research but also motivates new research directions focused on community-level leadership, nested organisational patterns, and the joint role of topology and attributes in shaping networked systems.

The central idea consists of redefining the concept of leadership, moving beyond a reductive notion based solely on high connectivity. Throughout the thesis, leaders emerge not only as high-degree nodes but as actors occupying structurally relevant positions, functioning as points of attraction or bridges, and contributing substantially to the cohesion of both groups and the system as a whole.

Across the different parts of the thesis, the notion of leadership is progressively refined by focusing on distinct yet complementary mechanisms of group formation. The analysis first distinguishes between structural and role-based equivalence, showing how leaders can be identified as nodes that perform similar functional roles within the system. This perspective highlights the importance of role-based patterns in shaping communities beyond neighbourhood overlap.

Leadership is then examined in relation to spatial density, illustrating how locally prominent units act as organising centres that guide spatial segmentation. In this

setting, prominence emerges from concentration rather than from connectivity alone, reinforcing the idea that leadership may be context-dependent and locally defined.

Building on these insights, the density-based framework identifies leaders as peaks of structural density. This formulation clarifies how attraction towards structurally prominent nodes drives community formation, while also revealing its limitations in contexts characterised by a moderate number of inter-community links, where structural prominence alone may fail to stabilise community boundaries.

The analysis further extends the definition of leadership by shifting attention from structural prominence to attribute-based community representativeness. In this context, representativeness is defined in terms of similarity to other group members in the attribute space, capturing how closely a node reflects the shared characteristics of its community. Leaders are thus identified as central nodes in the attribute space, showing how homophily can actively shape aggregation mechanisms and redefine influence in terms of similarity rather than popularity.

Finally, leadership is translated from the node level to the community level. Communities themselves are characterised according to the role they play within the system, distinguishing between core and peripheral roles and analysing how community leaders contribute to a hierarchical organisation of inter-community relations.

Overall, the thesis shows that degree heterogeneity and leader influence not only shape the internal structure of communities, but are also reflected in the organisation of relations between communities, giving rise to core–periphery patterns. In particular, the analysis highlights the role of leaders in peripheral communities, which seems to be crucial for integrating peripheral groups into the overall network.

This evidence is consistent with Granovetter (1973) contribution on the strength of weak ties, according to which network cohesion relies not only on strong intra-community ties, but crucially on weak ties that connect otherwise distinct clusters, enabling information flow across groups. Within this perspective, leadership is not only a property distinguishing core and peripheral community members, but also an internal feature of communities themselves: even peripheral groups exhibit hierarchical organisation, in which a small number of actors play a crucial role in sustaining connectivity. Peripheral leaders can be interpreted as actors who, through relatively weak but structurally strategic ties, enable the integration of otherwise isolated groups. In the absence of such nodes, many peripheral communities would remain marginal or disconnected, preventing the emergence of the core–periphery structure itself.

From a theoretical perspective, these results engage with the recent literature on the mechanisms underlying core–periphery formation, and in particular with the work of Ureña-Carrión et al. (2023). However, while that contribution analyses the emergence of the core at the node level as the outcome of the interaction between preferential attachment and homophily, this thesis shows how analogous dynamics may manifest at the community level.

It is important to emphasise that this work focuses on the analysis of observed networks, rather than on the direct simulation of formation processes. The results should therefore be interpreted as empirical and interpretative evidence, rather than as causal demonstrations of the underlying mechanisms.

Overall, the thesis suggests that a full understanding of network structures requires a joint analysis of micro-level mechanisms operating at the node level and macro-level configurations emerging at the community level. From this perspective, the work provides a methodological and conceptual foundation upon which integrated models can be built, capable of linking leadership, homophily, and hierarchical organisation within a unified and interpretable framework.

Appendix A

Open Science

This appendix documents the data sources, availability conditions, and computational resources used throughout the thesis, in line with open science principles and current best practices for transparency and reproducibility.

A.1 Data availability

Chapter 1 – Labour market networks The data used in Chapter 1 were provided by the *Regional Observatory on Policies and the Labour Market of the Friuli Venezia Giulia Region*. These data consist of administrative labour market records and were made available for research purposes under specific institutional agreements. Due to confidentiality constraints, the raw data cannot be publicly shared.

Chapter 2 – Housing market analysis The data analysed in Chapter 2 were provided within the framework of the ERC project *HABITAT*. Access to the underlying data is regulated by the project's data governance policy and is therefore restricted.

Chapter 4 – Horizon collaboration networks The data used in Chapter 4 are openly available through `data.europa.eu`, the official European data portal. In particular, project-level information was retrieved from the following datasets:

- Horizon 2020 projects
- Horizon Europe projects (2021–2027)

The processed network data generated for this thesis are publicly available under the Creative Commons Attribution 4.0 International License (CC BY 4.0) on *Zenodo*.

Repository title: Horizon Projects Network
DOI: `https://doi.org/10.5281/zenodo.13765372`

Chapter 5 – Co-authorship networks The data analysed in Chapter 5 were collected within the framework of the PRIN project *MASCONET*. The dataset was constructed by combining administrative records from the Italian Ministry of University and Research with bibliometric information extracted from Scopus. Due to licensing restrictions associated with bibliometric data, the raw data cannot be publicly disseminated but are available from the authors upon reasonable request.

A.2 Code availability

All computational analyses in this thesis were developed primarily using the R programming language, a widely used open-source environment for statistical computing and data analysis. Network analysis was conducted mainly using the igraph package (Csárdi and Nepusz, 2006). All analyses involving the extraction of embeddings—including semantic and network-based representations—were computed using Python, utilising its ecosystem of libraries for machine learning and natural language processing.

The development versions of the code used for this thesis are maintained using Git and hosted on https://github.com/sarageremia. At the time of writing, the project repositories remain private due to ongoing journal submissions. The code will be made publicly available upon acceptance of the related manuscripts and is currently available from the author upon request.

Whenever data access restrictions apply, the thesis provides detailed methodological descriptions to ensure reproducibility of the analytical workflow. For all chapters relying on open data, both data and code are documented in a manner that allows independent replication of the results.

Appendix B

Appendix to Chapter 2

B.1 Results for alternative clustering configurations

Table B.1 reports the predictive performance of the hedonic models when locational attributes are constructed using constrained K-means and spectral clustering under different average cluster sizes. These results illustrate the robustness of the estimates to the clustering specification and highlight that spectral clustering with $K = 303$ (approximately 300 flats per cluster) yields the strongest predictive performance.

B.2 Additional application: results for Hamburg

We conduct an external validation exercise using a second major German city: Hamburg. This analysis serves two purposes. First, it evaluates whether the proposed clustering framework generalizes to a different urban environment with distinct spatial and market characteristics. Second, by applying the method to another independently gridded dataset, it provides preliminary evidence on the robustness of the approach with respect to the placement of grid boundaries. Although only one additional city is presented here due to data-access constraints, the results nonetheless offer insight into how the method performs outside the Berlin context.
We analysed 20,843 flats for sale, aggregated to the same $1\,\text{km}^2$ spatial resolution

Locational Attribute	Mean Cluster Size	K	RMSE	R^2	MAE
K-Means Clustering	200	227	0.29	0.82	0.22
	300	303	0.29	0.81	0.22
	400	454	0.28	0.82	0.21
Spectral Clustering	200	227	0.29	0.81	0.22
	300	303	0.25	0.85	0.19
	400	454	0.29	0.81	0.22

TABLE B.1: Predictive performance of models using locational attributes derived from constrained K-means and spectral clustering across varying mean cluster sizes.

as in the main application. As in Berlin, spectral clustering is implemented with $K = 70$, corresponding to an average of approximately 300 flats per cluster. Table B.2 summarizes the predictive performance of the hedonic models under three locational specifications: no locational controls, constrained K-means clustering, and constrained spectral clustering.

Across all metrics, constrained spectral clustering again outperforms both the baseline model and the K-means specification. The reduction in prediction error (6.3% in RMSE and 7.3% in MAE) and the increase in explanatory power (2.2%) are consistent with the improvements observed in Berlin. However, the relative performance gains are slightly smaller than in the Berlin application. A plausible explanation is the lower heterogeneity in Hamburg's housing market and urban morphology—reflected in more homogeneous price levels and less spatial variability— which implies a higher baseline explanatory power and therefore less room for improvement. Overall, the results demonstrate that the method is not city-specific and that its advantages persist under a different spatial configuration.

Locational Attribute	Performance Metrics			% Change vs. Baseline		
	RMSE	R^2	MAE	ΔRMSE	ΔR^2	ΔMAE
None (Baseline)	0.26	0.85	0.20	–	–	–
K-Means Clustering	0.26	0.85	0.19	+2.7%	+1.0%	+3.2%
Spectral Clustering	0.25	0.87	0.19	+6.3%	+2.2%	+7.3%

TABLE B.2: Predictive performance of hedonic models for Hamburg using different locational attributes (K = 70), with percentage changes relative to the baseline model without locational controls.

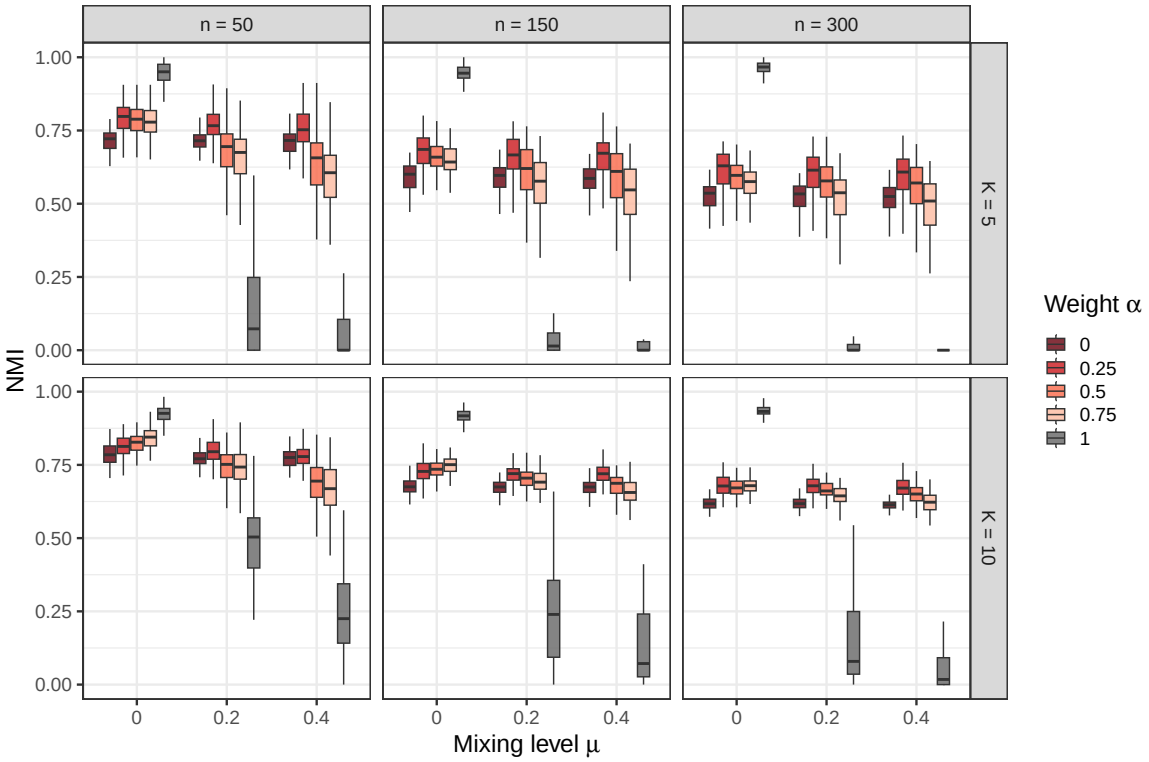


*FIGURE C1: **Adjacency matrix** – Normalized Mutual Information (NMI) distribution across different numbers of nodes n, numbers of communities K, mixing levels μ, and attribute–structure weights α. Results are shown for networks with non-uniform (Dirichlet-distributed) community sizes.*

Appendix C

Appendix to Chapter 3

C.1 Additional results of simulation study

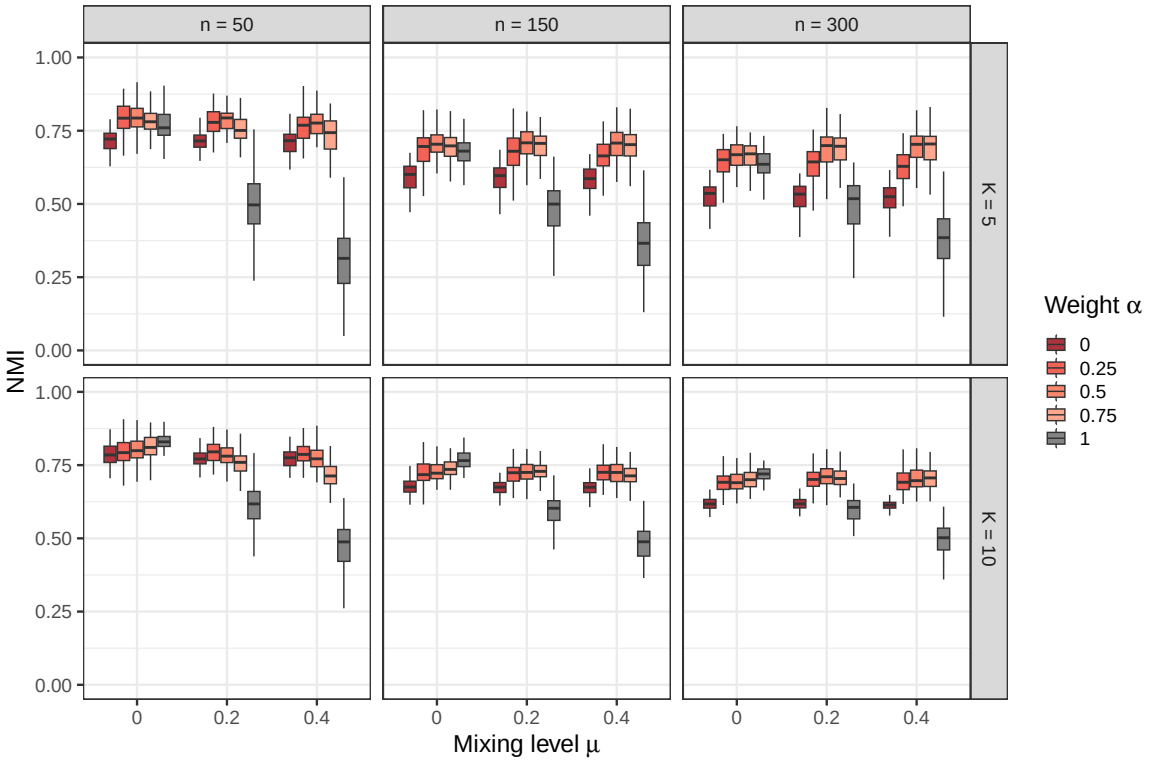


FIGURE C2: ***Jaccard matrix*** *– Normalized Mutual Information (NMI) distribution across different numbers of nodes n, numbers of communities K, mixing levels μ, and attribute–structure weights α. Results are shown for networks with non-uniform (Dirichlet-distributed) community sizes.*

Appendix D

Appendix to Chapter 4

D.1 Details of data generation in simulation study

We explore multiple network sizes, $n = \{50, 150, 300\}$, and numbers of communities, $K = \{5, 10\}$. Nodes are assigned to communities according to a probability distribution $\boldsymbol{\tau} = (\tau_1, \ldots, \tau_k, \ldots, \tau_K)$. This probability distribution is defined in two alternative ways:

- **Uniform**

  Each community is equally likely, hence

  $$\tau_k = \frac{1}{K}, \qquad k = 1, \ldots, K.$$

- **Non-uniform**

  The community proportions are drawn from a symmetric Dirichlet distribution:

  $$\boldsymbol{\tau} \sim \text{Dirichlet}(\mathbf{1}_K).$$

  Node labels are then drawn as

  $$q_i \mid \boldsymbol{\tau} \sim \text{Categorical}(\boldsymbol{\tau}).$$

Degree heterogeneity is introduced through node-specific parameters $\hat{\gamma}_i$, derived from the attribute densities produced by the GMM, with $\overline{\hat{\gamma}}$ being the mean attribute density. Let δ_i denote the density of node i derived from $\hat{\gamma}_i$. We first apply an exponential transformation:

$$\gamma_i^* = \exp\left(\frac{\hat{\gamma}_i}{\overline{\hat{\gamma}}}\right),$$

which amplifies the differences between high- and low-density nodes. We then derive node density δ_i rescaling the transformed γ_i^* to the $[0, 1]$ interval:

$$\delta_i = \frac{\gamma_i^* - \min(\boldsymbol{\gamma}^*)}{\max(\boldsymbol{\gamma}^*) - \min(\boldsymbol{\gamma}^*)}.$$

To prevent low-density nodes from becoming isolated, we impose a minimum within-community bound on degree parameters. To maintain controlled sparsity, within-community edge probabilities scale with $\log(n)/n$, ensuring that, as $n \to \infty$, the expected degree remains of order $\mathcal{O}(\log n)$ to preserve both community detectability and connectivity in sparse network models. We set

$$\rho_{\min} = 5 \cdot \frac{\log n}{n},$$

and then rescale

$$\delta_i \leftarrow \delta_i(1 - \rho_{\min}) + \rho_{\min},$$

which guarantees $\delta_i \in [\rho_{\min}, 1]$ and yields realistic leader–follower heterogeneity while preserving sparsity.

Network edges are generated via a DC–SBM. Within community q, the probability that nodes i and j form a link is proportional to

$$\Pr(i \leftrightarrow j \mid C(v_i) = C(v_j) = q) \propto \delta_i \delta_j.$$

Nodes with larger δ therefore act as local leaders, attracting more intra-community ties.

In summary, this simulation design combines:

1. **Attribute-driven heterogeneity:** nodes located in dense regions of attribute space become leaders with larger δ_i.
2. **Sparse community structure:** connectivity probabilities scale like $\log(n)/n$, producing realistic networks as size increases.
3. **Flexible cluster imbalance:** community sizes can be uniform or drawn from a Dirichlet distribution.

The resulting networks exhibit community structure driven jointly by discrete block membership and continuous attribute-based attraction, providing a realistic setting to evaluate whether the proposed AttDeCoDe algorithm can recover the true partition.

D.2 Additional results of simulation study

Figure D1 reports additional results for the simulation study of Section 4.3.1. Across most configurations, the GMM and kNN density specifications improve cluster detectability compared to the original DeCoDe, consistent with the AttDeCoDe results discussed in the main text.

Figure D2 reports NMI values for DC-SBM networks with non-uniform community sizes. Overall, the patterns closely mirror those observed in the equal-size setting

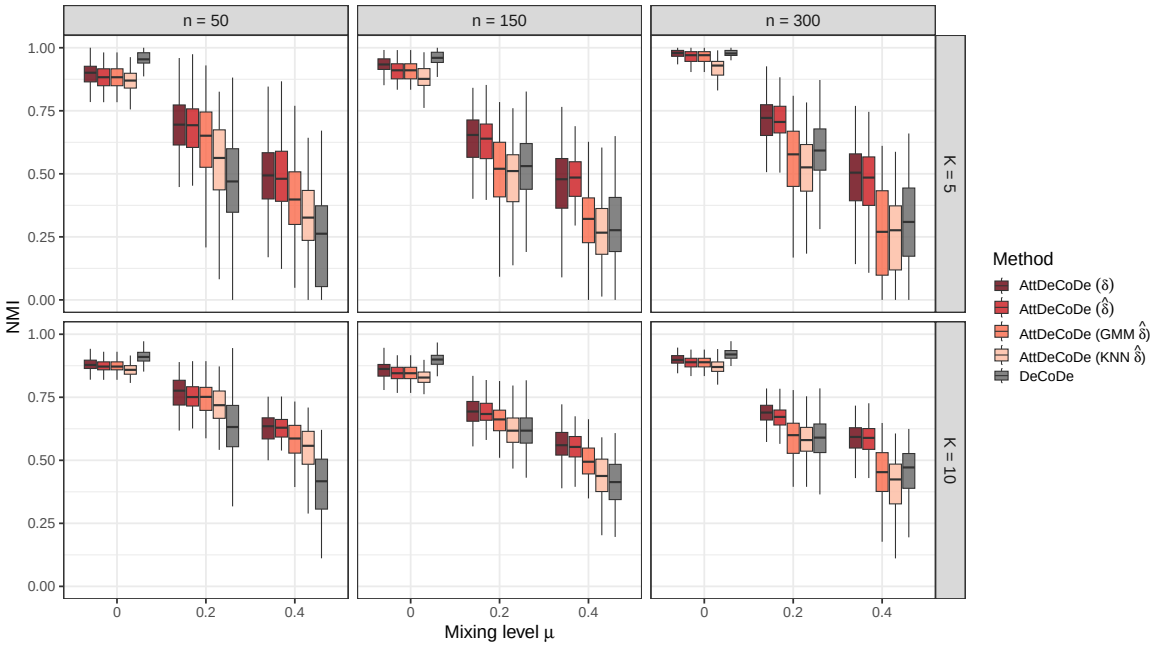


*FIGURE D1: Normalised Mutual Information (NMI) distribution across different network sizes and numbers of communities (K) for **uniform** community sizes. Results are shown for AttDeCoDe (all density estimators) and DeCoDe.*

discussed in the main text, with the proposed AttDeCoDe variants consistently achieving high levels of agreement with the true partition. The primary difference lies in the increased variability of the results, which is particularly pronounced for the baseline methods.

Figure D3 reports NMI values for networks with non-uniform community sizes, comparing AttDeCoDe under different density estimators with the original DeCoDe. Relative to the results shown in Figure D1, all methods exhibit greater variability and lower overall performance. This behaviour is consistent with the patterns observed in Figure D2 and reflects the increased complexity of the setting, which allows for the presence of very small communities.

Figure D4 reports ARI values for DC-SBM networks with uniform community sizes. Overall, the patterns closely mirror those observed in NMI setting discussed in the main text.

Figure D5 reports ARI values for networks with non-uniform community sizes, showing patterns consistent with the NMI results discussed above.

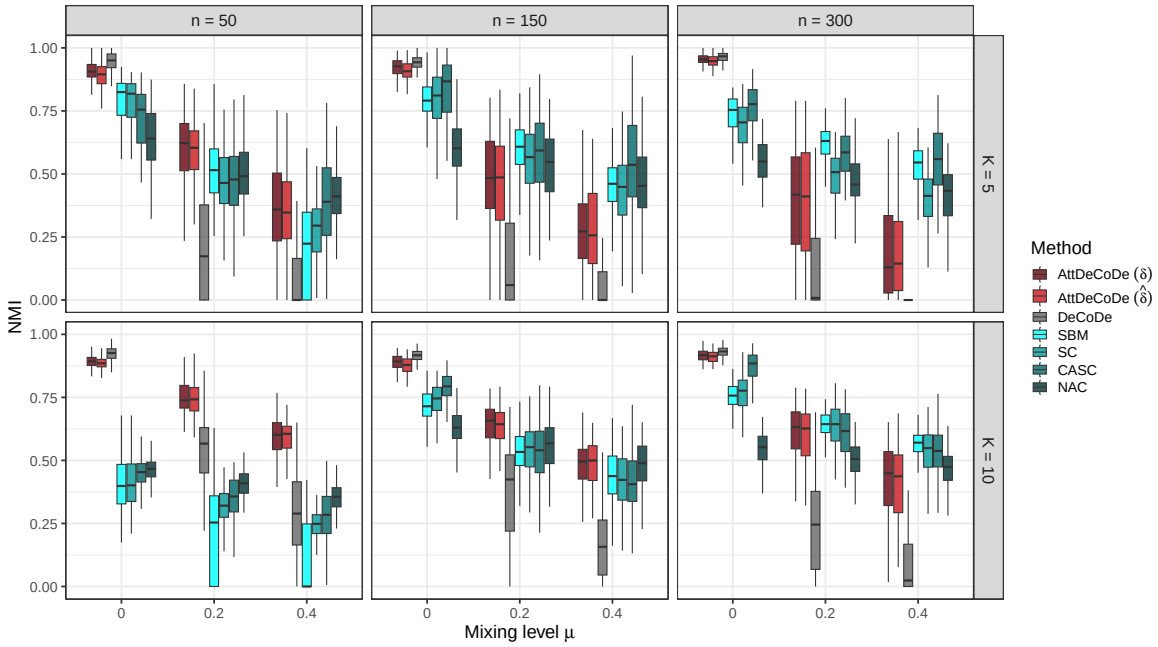


*FIGURE D2: Normalised Mutual Information (NMI) distribution across different network sizes and numbers of communities (K) for **non-uniform** community sizes. Results are reported for all competing methods, including AttDeCoDe, DeCoDe, binary stochastic block model (SBM), Spectral Clustering (SC), Covariate-Assisted Spectral Clustering (CASC), and SC on Network-Adjusted Covariates (NAC).*

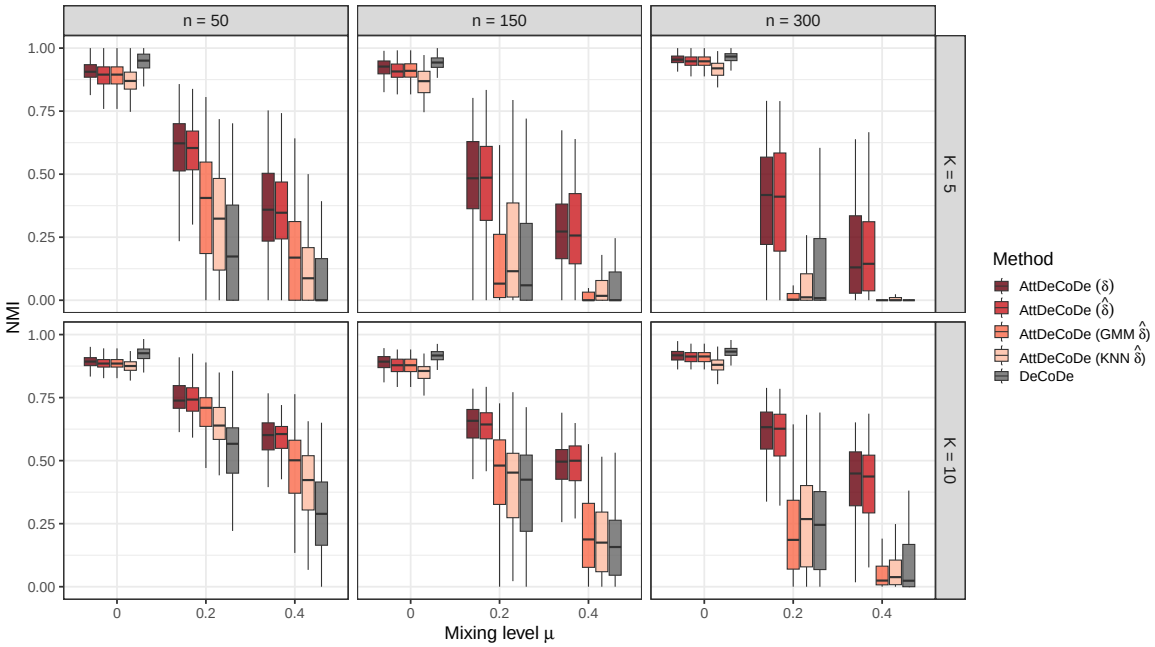


*FIGURE D3: Normalised Mutual Information (NMI) distribution across different network sizes and numbers of communities (K) for **non-uniform** community sizes. Results are shown for AttDeCoDe (all density estimators) and DeCoDe.*

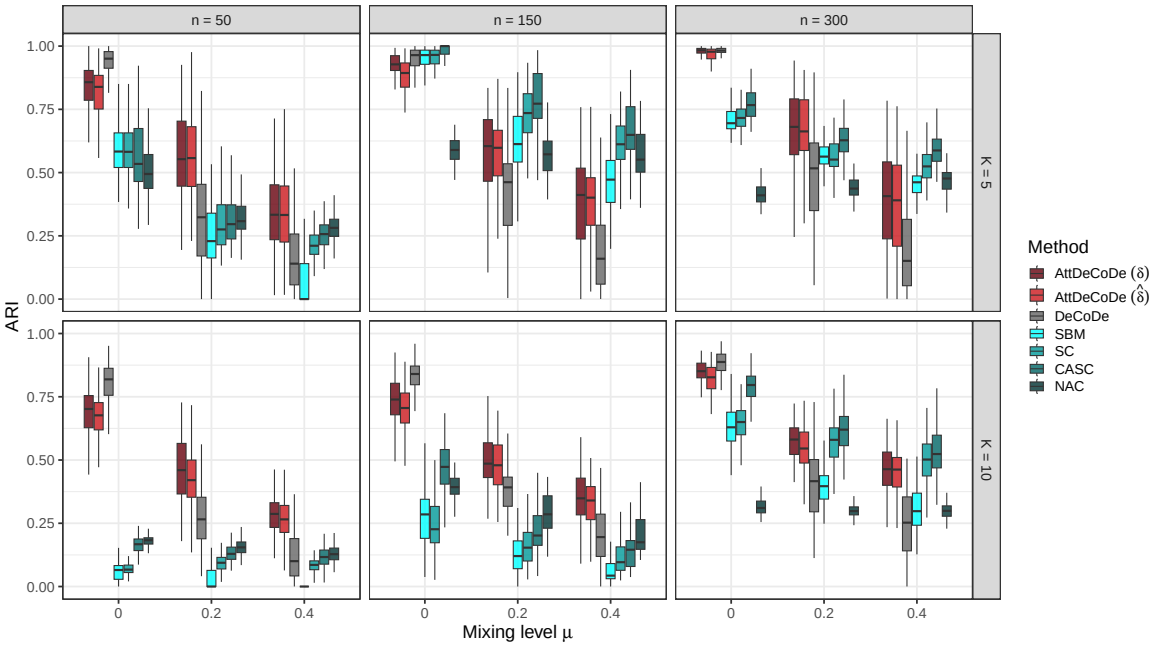


*FIGURE D4: Adjusted Rand Index (ARI) distribution across different network sizes and numbers of communities (K) for **uniform** community sizes. Results are reported for all competing methods, including AttDeCoDe, DeCoDe, binary stochastic block model (SBM), Spectral Clustering (SC), Covariate-Assisted Spectral Clustering (CASC), and SC on Network-Adjusted Covariates (NAC).*

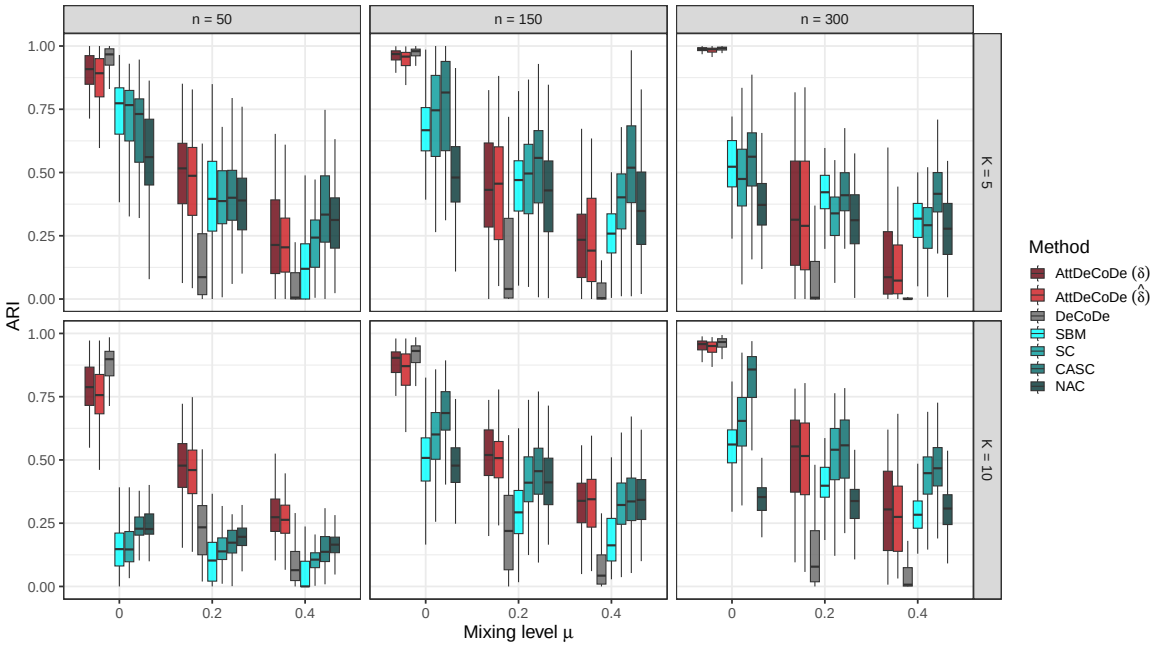


*FIGURE D5: Adjusted Rand Index (ARI) distribution across different network sizes and numbers of communities (K) for **non-uniform** community sizes. Results are reported for all competing methods, including AttDeCoDe, DeCoDe, binary stochastic block model (SBM), Spectral Clustering (SC), Covariate-Assisted Spectral Clustering (CASC), and SC on Network-Adjusted Covariates (NAC).*

Appendix E

Appendix to Chapter 5

E.1 Details of data generation in simulation study

Different network sizes, $n = \{100, 200, 500, 1000\}$ and numbers of communities, $K = \{5, 10, 15\}$ are considered. Each node is assigned to one of these communities according to a probability distribution $\boldsymbol{\tau} = (\tau_1, \ldots, \tau_k, \ldots, \tau_K)$:

- **Uniform**

  Each community is equally likely, so

  $$\tau_k = \frac{1}{K}, \qquad k = 1, \ldots, K,$$

- **Non-uniform**

  The community proportions are drawn from a symmetric Dirichlet distribution:

  $$\boldsymbol{\tau} \sim \text{Dirichlet}(\mathbf{1}_K).$$

Node labels are then drawn as

$$q_i \mid \boldsymbol{\tau} \sim \text{Categorical}(\boldsymbol{\tau}).$$

A parameter $\lambda = \{25, 50, 75\}$ controls the percentage of communities belonging to the core.

Networks are generated according to a SBM whose connectivity probability parameters are defined as a function of n (number of nodes) and community and core-periphery membership. Some considerations concerning the definition of these connectivity probabilities. In sparse SBMs, considering an expected node degree of at least $\log(n)$ can be relevant for ensuring connectivity and community structure detection, particularly when connection probabilities scale as $\log(n)/n$. In line with this theory, all our connection probabilities are a function of $\log(n)/n$, so that the graph remains sparse as $n \to \infty$. These probabilities are adjusted by a scaling factor (tuned by means of empirical evaluation and sensitivity analysis) which controls the relative density of connections between different types of node pairs (Table E1), allowing the generation of networks with both community and core-periphery structures.

E.2 Alternative summary statistics

An alternative specification of the objective function in Equation 5.1 replaces the mean edge frequencies with the median edge frequencies of the core and periphery, denoted with m_c and m_p and replaces the standard deviations with the interquartile ranges IQR_c and IQR_p. The resulting alternative objective function is

$$\tilde{\phi}(\mathbf{z}; \hat{\mathbf{\Theta}}) = \big(\delta_c + m_c - IQR_c\big) - \big(\delta_p + m_p + IQR_p\big). \tag{E.1}$$

Figure E1 reports the values of $\tilde{\phi}(\mathbf{z}; \hat{\mathbf{\Theta}})$ for the simulation study of Section 5.3.2. While using more robust statistics such as medians and interquartile ranges may improve robustness to skewed or asymmetric connection densities, their discrete nature can create flat regions in the objective surface and generate multiple local optima. As a result, optimization becomes less stable, potentially complicating the identification of a unique community-level core–periphery partition.

E.3 Additional results of simulation study

Figure E2 reports F1 score values for the simulation study of Section 5.3.3. Results with regards to F1 scores are in line with the BA results commented in the main text.

Figure E3 shows the estimated number of clusters for each node-level community detection method (Louvain, Infomap, and SBM) in the simulation study of Section 5.3.3. In scenarios where the data-generating value of K is large, all three methods struggle to recover the true number of communities: Louvain and SBM typically underestimate the number of clusters, whereas Infomap tends to overestimate it. Despite these inaccuracies, the results for core–periphery identification remain satisfactory (see Figures 5.3 and 5.4 in the main text). In particular, although the node-level SBM frequently underestimates K, the method nevertheless identifies the core–periphery partition correctly and consistently. This suggests that, when the number of clusters is underestimated, the estimated node-level communities tend to merge true communities within the core, preserving the higher-order core–periphery structure even if the finer-grained clustering is not fully recovered.

Edge type	**Connection probability**
Within-community	$8 \cdot \log(n)/n$
Core–core	$\log(n)/n$
Core–periphery	$(1/8) \cdot \log(n)/n$
Periphery–periphery	$(1/40) \cdot \log(n)/n$

TABLE E1: Scaling of edge probabilities by edge type.

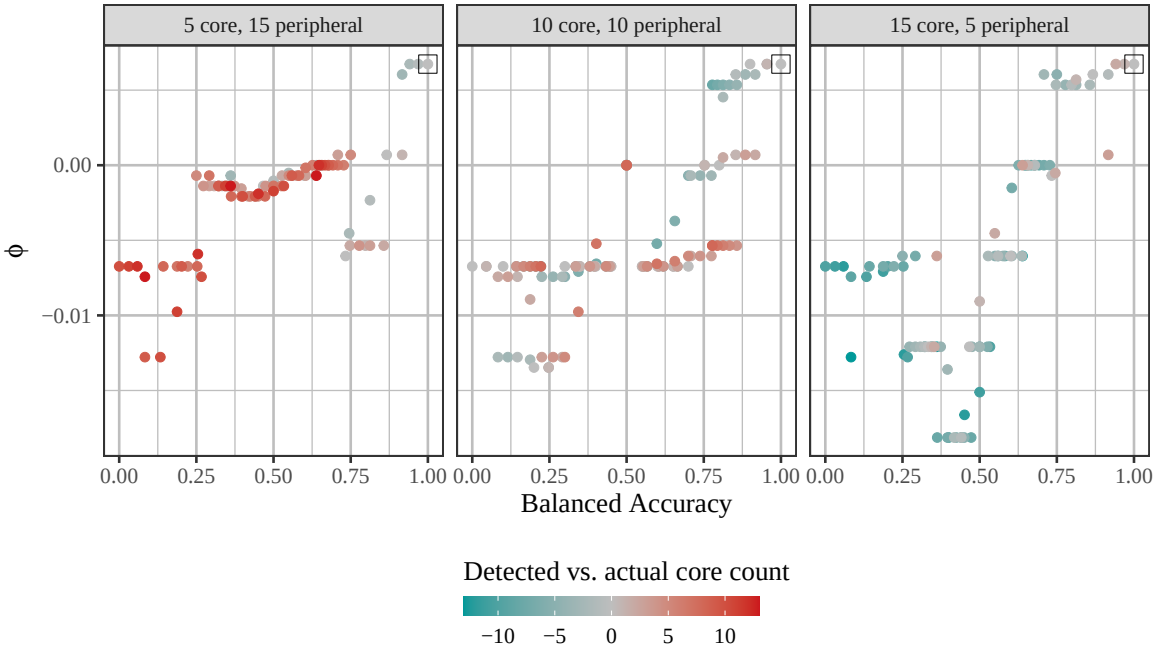
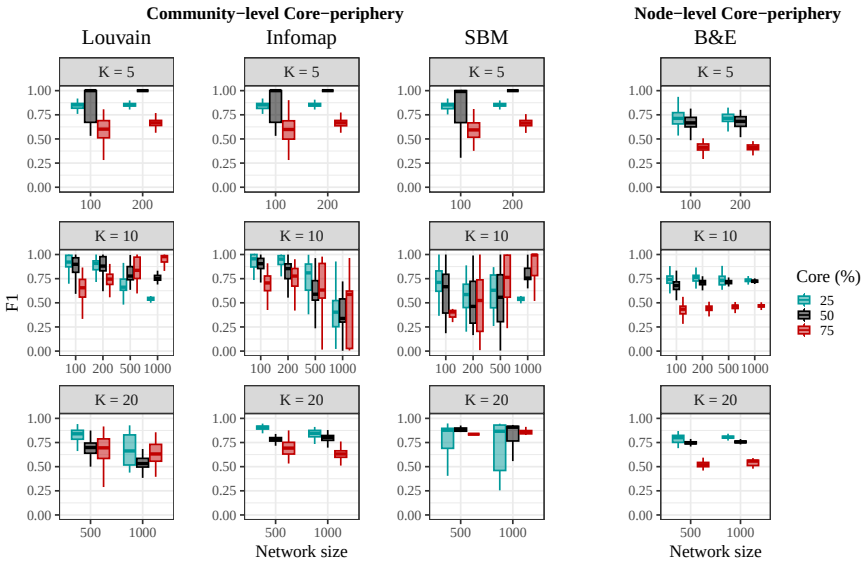


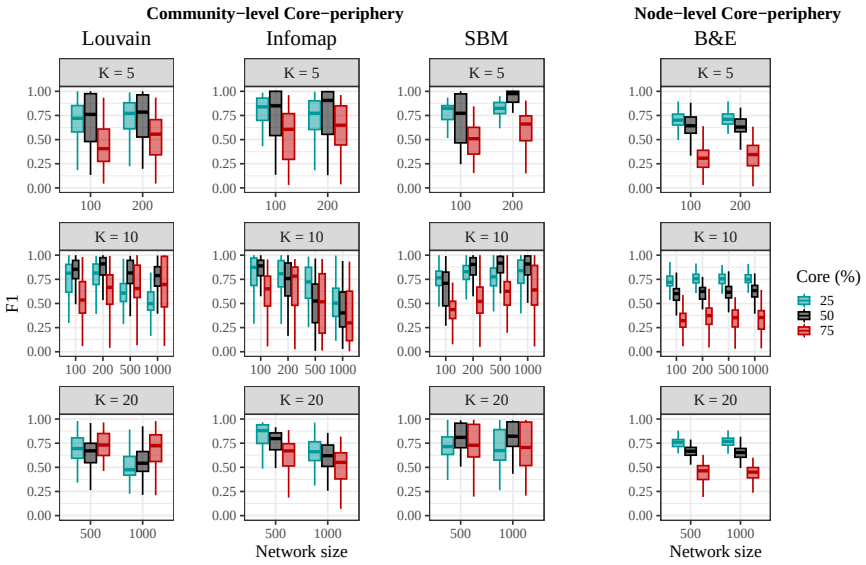
FIGURE E1: Scatter plot of the alternative objective function $\tilde{\phi}(\mathbf{z}; \hat{\boldsymbol{\Theta}})$ using median and IQR vs. balanced accuracy for core-periphery networks with $n = 1000$ nodes, $K = 20$ communities, and core proportions of 25%, 50%, and 75%. Blue points represent solutions with more peripheral communities than the true structure, while red points indicate solutions with more core communities. The true solution (Balanced Accuracy = 1) is highlighted with a square.

E.4 tf-idf distributions

Figure E4 reports the topic tf-idf and title tf-idf distributions for the co-authorship network of Section 5.4; see the main text for discussion.

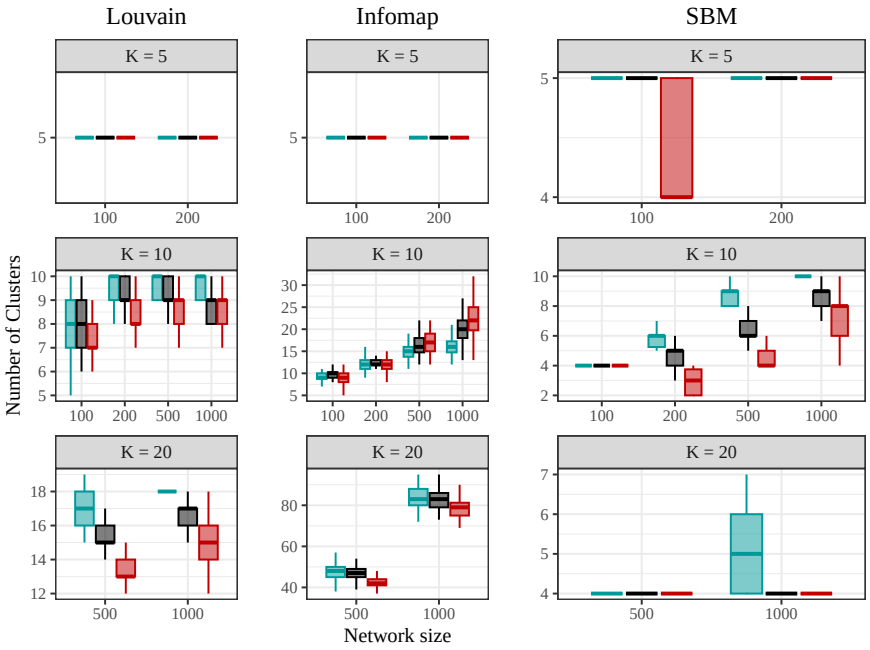


(A) Uniform

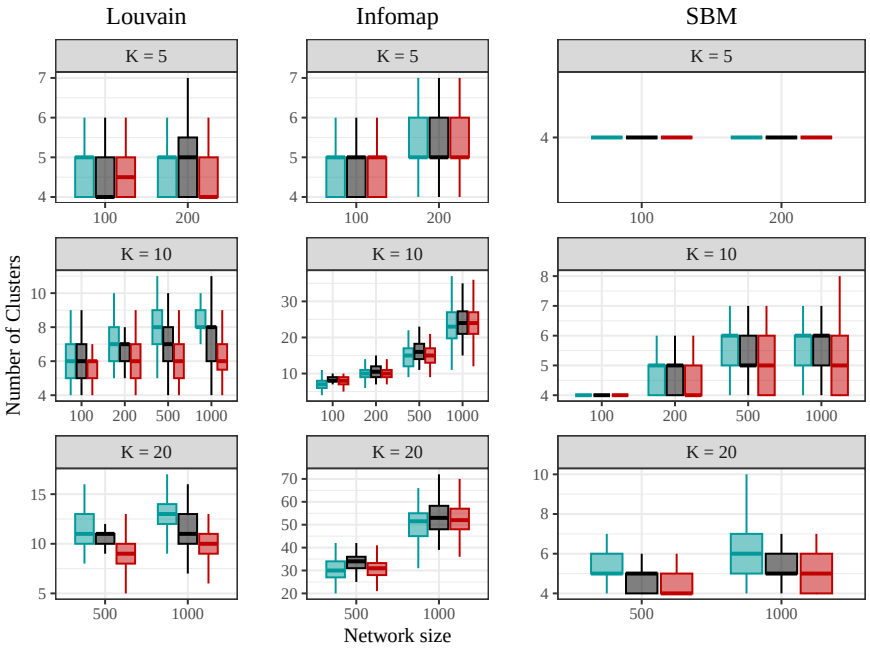


(B) Non-uniform

FIGURE E2: (a) F1 score distribution across different network sizes (n) and numbers of communities (K) for uniform community sizes. (b) F1 score distribution across different network sizes (n) and numbers of communities (K) for non-uniform (Dirichlet-distributed) community sizes.

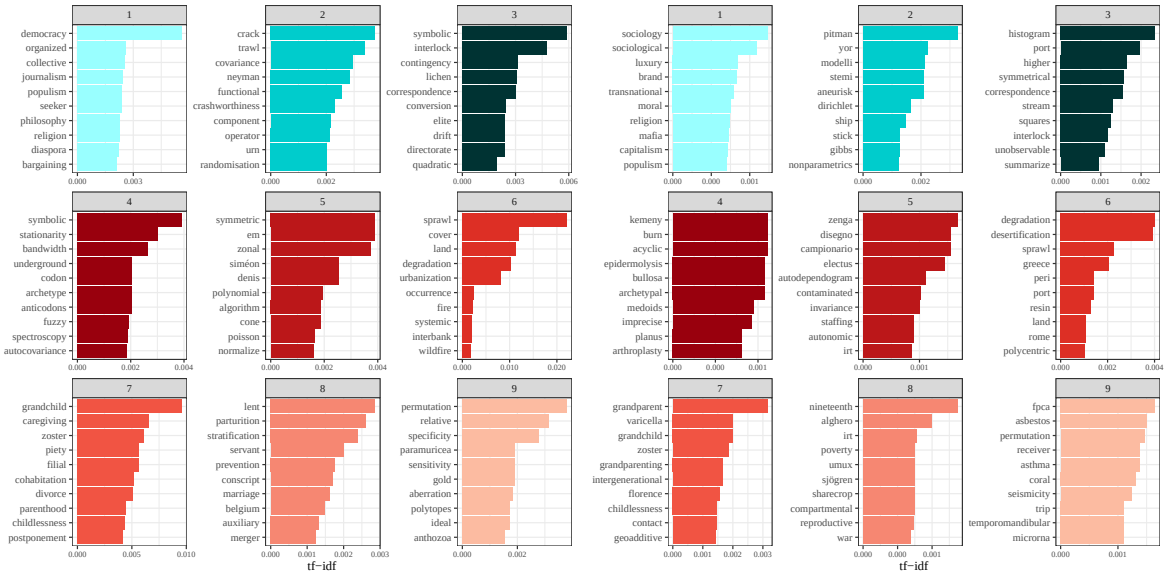


(A) Uniform



(B) Non-uniform

FIGURE E3: (a) Estimated number of clusters distribution across different network sizes (n) and numbers of communities (K) for uniform community sizes. (b) Estimated number of clusters distribution across different network sizes (n) and numbers of communities (K) for non-uniform (Dirichlet-distributed) community sizes.



(A) Paper topic *(B) Paper title*

FIGURE E4: (a) Distribution of topic tf-idf by cluster and core-periphery organisation. (b) Distribution of title tf-idf by cluster and core-periphery organisation.

Bibliography

Ahajjam, S., El Haddad, M., and Badir, H. (2018). A new scalable leader-community detection approach for community detection in social networks. *Social Networks*, 54:41–49.

Ahmed, N. K., Rossi, R., Boaz Lee, J., Willke, T. L., Zhou, R., Kong, X., and Eldardiry, H. (2018). Learning Role-based Graph Embeddings. *arXiv e-prints*, page arXiv:1802.02896.

Akachar, E., Bougteb, Y., Ouhbi, B., and Frikh, B. (2025). LeaDCD: Leadership concept-based method for community detection in social networks. *Information Sciences*, 686:121341.

Alba, R. D. and Kadushin, C. (1976). The intersection of social circles. A new measure of social proximity in networks. *Sociological Methods and Research*, 5:77 – 102.

Alba, R. D. and Moore, G. (1978). Elite social circles. *Sociological Methods & Research*, 7(2):167–188.

Alfò, M., Nieddu, L., and Vicari, D. (2008). A finite mixture model for image segmentation. *Statistics and Computing*, 18(2):137–150.

Alfó, M., Nieddu, L., and Vicari, D. (2009). Finite mixture models for mapping spatially dependent disease counts. *Biometrical Journal: Journal of Mathematical Methods in Biosciences*, 51(1):84–97.

Anselin, L. (1988). *Spatial econometrics: methods and models*, volume 4. Springer Science & Business Media.

Anselin, L. (2024). *An introduction to spatial data science with GeoDa: Volume 1: Exploring spatial data*. Chapman and Hall/CRC.

Anselin, L., Amaral, P., Anselin, L., and Amaral, P. (2023). Endogenous spatial regimes. *Journal of Geographical Systems 2023 26:2*, 26:209–234.

Arioli, R., Bobeica, E., Roma, M., and Soudan, M. (2023). Rent inflation in the euro area. *Economic Bulletin Boxes*, 7.

Arlia, D. (2024). Labor Market Shocks across Heterogeneous Housing Markets.

Asikainen, A., Iñiguez, G., Ureña-Carrión, J., Kaski, K., and Kivelä, M. (2020). Cumulative effects of triadic closure and homophily in social networks. *Science Advances*, 6(19):eaax7310.

Assunção, R. M., Neves, M. C., Câmara, G., and Freitas, C. D. C. (2006). Efficient regionalization techniques for socio-economic geographical units using minimum spanning trees. *International Journal of Geographical Information Science*, 20:797–811.

Avery, C. and Pathak, P. A. (2021). The distributional consequences of public school choice. *American Economic Review*, 111(1):129–152.

Bacci, S., Bertaccini, B., and Petrucci, A. (2023). Insights from the co-authorship network of the italian academic statisticians. *Scientometrics*, 128(8):4269–4303.

Bachmann, J., Martin-Gutierrez, S., Espín-Noboa, L., Cinardi, N., and Karimi, F. (2025). Network inequality through preferential attachment, triadic closure, and homophily. *arXiv preprint arXiv:2509.23205*.

Balland, P.-A., Boschma, R., Crespo, J., and Rigby, D. L. (2019a). Smart specialization policy in the European Union: Relatedness, knowledge complexity and regional diversification. *Regional Studies*, 53(9):1252–1268.

Balland, P.-A., Boschma, R., and Ravet, J. (2019b). Network dynamics in collaborative research in the EU, 2003–2017. *European Planning Studies*, 27(9):1811–1837.

Barabási, A. L. and Albert, R. (1999). Emergence of scaling in random networks. *Science*, 286:509–512.

Barber, M. J., Krueger, A., Krueger, T., and Roediger-Schluga, T. (2006). Network of European Union–funded collaborative research and development projects. *Physical Review E*, 73(3):036132.

Barber, M. J. and Scherngell, T. (2013). The community structure of european r&d collaboration. In *The Geography of Networks and R&D Collaborations*, pages 151–173. Springer.

Bassett, D. S., Wymbs, N. F., Rombach, M. P., Porter, M. A., Mucha, P. J., and Grafton, S. T. (2013). Task-based core-periphery organization of human brain dynamics. *PLOS Computational Biology*, 9(9):e1003171.

Batagelj, V. and Zaveršnik, M. (2003). An O(m) algorithm for cores decomposition of networks. *arXiv:cs/0310049*.

Bayer, P., Ferreira, F., and McMillan, R. (2007). A unified framework for measuring preferences for schools and neighborhoods. *Journal of political economy*, 115(4):588–638.

Beygelzimer, A., Kakadet, S., Langford, J., Arya, S., Mount, D., and Li, S. (2024). *FNN: Fast Nearest Neighbor Search Algorithms and Applications*. R package version 1.1.4.1.

Bianconi, G. and Barabási, A. L. (2001). Competition and multiscaling in evolving networks. *Europhysics Letters*, 54:436.

Bianconi, G., Darst, R. K., Iacovacci, J., and Fortunato, S. (2014). Triadic closure as a basic generating mechanism of the structure of complex networks. *Physical review. E, Statistical, nonlinear, and soft matter physics*, 90 4:042806.

Binkiewicz, N., Vogelstein, J. T., and Rohe, K. (2017). Covariate-assisted spectral clustering. *Biometrika*, 104(2):361–377. Publisher Copyright: © 2017 Biometrika Trust.

Borgatti, S. P. and Everett, M. G. (2000). Models of core/periphery structures. *Social Networks*, 21(4):375–395.

Borgoni, R., Michelangeli, A., and Pontarollo, N. (2018). The value of culture to urban housing markets. *Regional Studies*, 52(12):1672–1683.

Bouveyron, C., Celeux, G., Murphy, T. B., and Raftery, A. E. (2019). *Model-Based Clustering and Classification for Data Science: With Applications in R*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press.

Boyd, J. P., Fitzgerald, W. J., and Beck, R. J. (2006). Computing core/periphery structures and permutation tests for social relations data. *Social Networks*, 28(2):165–178.

Brasington, D. M. and Hite, D. (2005). Demand for environmental quality: a spatial hedonic analysis. *Regional science and urban economics*, 35(1):57–82.

Breiger, R. L., Boorman, S. A., and Arabie, P. (1975). An algorithm for clustering relational data with applications to social network analysis and comparison with multidimensional scaling. *Journal of Mathematical Psychology*, 12(3):328–383.

Breschi, S. and Cusmano, L. (2004). Unveiling the texture of a European Research Area: emergence of oligarchic networks under EU Framework Programmes. *International Journal of Technology Management*, 27(8):747.

Brusco, M. (2011). An exact algorithm for a core/periphery bipartitioning problem. *Social Networks*, 33(1):12–19.

Cai, H., Zheng, V. W., and Chang, K. C.-C. (2018). A comprehensive survey of graph embedding: Problems, techniques, and applications. *IEEE transactions on knowledge and data engineering*, 30(9):1616–1637.

Calka, B. (2019). Estimating residential property values on the basis of clustering and geostatistics. *Geosciences*, 9(3):143.

Cerqueti, R., Iovanella, A., and Mattera, R. (2024). Clustering networked funded european research activities through rank-size laws. *Annals of Operations Research*, 342(3):1707–1735.

Chunaev, P. (2020). Community detection in node-attributed social networks: A survey. *Computer Science Review*, 37:100286.

Clinchant, S. and Perronnin, F. (2013). Aggregating continuous word embeddings for information retrieval. In Allauzen, A., Larochelle, H., Manning, C., and Socher, R., editors, *Proceedings of the Workshop on Continuous Vector Space Models and their Compositionality*, pages 100–109, Sofia, Bulgaria. Association for Computational Linguistics.

Cohen, E. (2019). node2vec: Python implementation. https://github.com/eliorc/node2vec. Python package.

Colizza, V., Flammini, A., Serrano, M. A., and Vespignani, A. (2006). Detecting rich-club ordering in complex networks. *Nature Physics*, 2(2):110–115.

Côme, E., Jouvin, N., Latouche, P., and Bouveyron, C. (2021). Hierarchical clustering with discrete latent variable models and the integrated classification likelihood. *Advances in Data Analysis and Classification*, 15(4):957–986.

Côme, E. and Latouche, P. (2015). Model selection and clustering in stochastic block models based on the exact integrated complete data likelihood. *Statistical Modelling*, 15(6):564–589.

Csárdi, G. and Nepusz, T. (2006). The igraph software package for complex network research. *InterJournal*, Complex Systems:1695.

Csárdi, G., Nepusz, T., Traag, V., Horvát, S., Zanini, F., Noom, D., and Müller, K. (2025). *igraph: Network Analysis and Visualization in R*. R package version 2.1.4.

Cucuringu, M., Rombach, P., Lee, S. H., and Porter, M. A. (2016). Detection of core–periphery structure in networks using spectral methods and geodesic paths. *European Journal of Applied Mathematics*, 27(6):846–887.

Cugmas, M., Ferligoj, A., and Kronegger, L. (2015). The stability of co-authorship structures. *Scientometrics 2015 106:1*, 106:163–186.

Cui, P., Wang, X., Pei, J., and Zhu, W. (2018). A survey on network embedding. *IEEE transactions on knowledge and data engineering*, 31(5):833–852.

Cui, P., Wu, L., Pei, J., Zhao, L., and Wang, X. (2022). Graph Representation Learning. *Graph Neural Networks: Foundations, Frontiers, and Applications*, 14(3):17–26.

Danon, L., Díaz-Guilera, A., Duch, J., and Arenas, A. (2005). Comparing community structure identification. *Journal of Statistical Mechanics: Theory and Experiment*, 2005(09):P09008–P09008.

De Stefano, D., Fabbrucci Barbagli, A. G., Santelli, F., and Zaccarin, S. (2023a). Collaboration networks: methodological issues and updated empirical evidence on Italian statisticians. In *IES2023 - Statistical Methods For Evaluation And Quality: Techniques, Technologies And Trends*. Edizioni Il Viandante.

De Stefano, D., Fuccella, V., Vitale, M. P., and Zaccarin, S. (2013). The use of different data sources in the analysis of co-authorship networks and scientific performance. *Social Networks*, 35(3):370–381.

De Stefano, D., Fuccella, V., Vitale, M. P., and Zaccarin, S. (2023b). Quality issues in co-authorship data of a national scientific community. *Network Science*, 11(1):98–112.

De Stefano, D., Giordano, G., and Vitale, M. P. (2011). Issues in the analysis of co-authorship networks. *Quality & Quantity*, 45(5):1091–1107.

De Stefano, D. and Zaccarin, S. (2013). Modelling Multiple Interactions in Science and Technology Networks. *Industry and Innovation*, 20(3):221–240.

Donthu, N., Kumar, S., Mukherjee, D., Pandey, N., and Lim, W. M. (2021). How to conduct a bibliometric analysis: An overview and guidelines. *Journal of Business Research*, 133:285–296.

Doreian, P., Batagelj, V., Ferligoj, A., and Block, G. (2007). P. Doreian, V. Batagelj, and A. Ferligoj, Generalized Block-modeling, Cambridge University Press, 2005, pp. xv + 384, ISBN 0521840856. *Journal of Classification*, 24(2):308–311.

Dubin, R. A. (1992). Spatial autocorrelation and neighborhood quality. *Regional science and urban economics*, 22(3):433–452.

Eriksson, R. H. (2011). Localized Spillovers and Knowledge Flows: How Does Proximity Influence the Performance of Plants? *Economic Geography*, 87(2):127–152.

Ester, M., Kriegel, H.-P., Sander, J., and Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, KDD'96, page 226–231. AAAI Press.

Estévez, J. L. and Nordlund, C. (2025). Revising the Borgatti-Everett core-periphery model: Inter-categorical density blocks and partially connected cores. *Social Networks*, 81:31–51.

Etzkowitz, H. and Leydesdorff, L. (1995). The triple helix–university-industry-government relations: A laboratory for knowledge based economic development. *EASST review*, 14(1):14–19.

European Commission (2024). Cordis: EU research results.

European Commission, for Research, D.-G., Innovation, Hollanders, H., and Es-Sadki, N. (2023). *Regional Innovation Scoreboard 2023*. Publications Office of the European Union.

European Union (2009). Commission recommendation of 29 October 2009 on the use of the International Standard Classification of Occupations (ISCO-08).

Fabbrucci Barbagli, A. G., De Stefano, D., Santelli, F., and Zaccarin, S. (2025). Unveiling collaboration persistence and interactions among italian academic statisticians through relational hyperevent models. *Statistical Methods & Applications*, pages 1–19.

Falkowski, T., Barth, A., and Spiliopoulou, M. (2007). DENGRAPH: A Density-based Community Detection Algorithm. In *IEEE/WIC/ACM International Conference on Web Intelligence (WI'07)*, pages 112–115. IEEE.

Feng, H. and Lu, M. (2013). School quality and housing prices: Empirical evidence from a natural experiment in shanghai, china. *Journal of Housing Economics*, 22(4):291–307.

Ferligoj, A. and Batagelj, V. (1982). Clustering with relational constraint. *Psychometrika*, 47(4):413–426.

Fortunato, S. (2010). Community detection in graphs. *Physics Reports*, 486(3-5):75–174.

Fortunato, S., Bergstrom, C. T., Börner, K., Evans, J. A., Helbing, D., Milojević, S., Petersen, A. M., Radicchi, F., Sinatra, R., Uzzi, B., et al. (2018). Science of science. *Science*, 359(6379):eaao0185.

Fortunato, S. and Hric, D. (2016). Community detection in networks: A user guide. *Physics Reports*, 659:1–44.

Fuhse, J. A. and Gondal, N. (2024). Networks from culture: Mechanisms of tie-formation follow institutionalized rules in social fields. *Social Networks*, 77:43–54. Network Ecology.

Gallagher, R. J., Young, J.-G., and Welles, B. F. (2021). A clarified typology of core-periphery structure in networks. *Science Advances*, 7(12).

Gallin, J. (2008). The long-run relationship between house prices and rents. *Real Estate Economics*, 36(4):635–658.

Galtung, J. (1971). A structural theory of imperialism. *Journal of Peace Research*, 8(2):81–117.

Ganganath, N., Cheng, C.-T., and Tse, C. K. (2014). Data Clustering with Cluster Size Constraints Using a Modified K-Means Algorithm. In *2014 International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery*, pages 158–161. IEEE.

Genova, V. G., Giordano, G., Ragozini, G., and Vitale, M. P. (2024). An analytic strategy for data processing of multimode networks. *Advances in Data Analysis and Classification*, 18(3):745–767.

Gevrey, M., Dimopoulos, Y., and Lek, S. (2003). Review and comparison of methods to study the contribution of variables in artificial neural networks models. *Ecological Modelling*, 160:249–264.

Gilderbloom, J. I., Hanka, M. J., and Ambrosius, J. D. (2012). Without bias? government policy that creates fair and equitable property tax assessments. *The American Review of Public Administration*, 42(5):591–605.

Goltsev, A. V., Dorogovtsev, S. N., and Mendes, J. F. (2006). K -core (bootstrap) percolation on complex networks: Critical phenomena and nonlocal effects. *Physical Review E - Statistical, Nonlinear, and Soft Matter Physics*, 73(5):056101.

Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep Learning*. MIT Press. http://www.deeplearningbook.org.

Gower, J. C. (1971). A general coefficient of similarity and some of its properties. *Biometrics*, pages 857–871.

Granovetter, M. S. (1973). The strength of weak ties. *American journal of sociology*, 78(6):1360–1380.

Grover, A. and Leskovec, J. (2016). Node2vec: Scalable feature learning for networks. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 13-17-Augu:855–864.

Guerrero, O. A. and Axtell, R. L. (2013). Employment Growth through Labor Flow Networks. *PLoS ONE*, 8(5):e60808.

Haurin, D. R. and Brasington, D. (1996). School quality and real house prices: Inter-and intrametropolitan effects. *Journal of Housing economics*, 5(4):351–368.

Hébert-Dufresne, L., Grochow, J. A., and Allard, A. (2016). Multi-scale structure and topological anomaly detection via a new network statistic: The onion decomposition. *Scientific Reports*, 6(1):1–9.

Helal, N. A., Ismail, R. M., Badr, N. L., and Mostafa, M. G. M. (2016). An Efficient Algorithm for Community Detection in Attributed Social Networks. In *Proceedings of the 10th International Conference on Informatics and Systems*, pages 180–184, New York, NY, USA. ACM.

Helal, N. A., Ismail, R. M., Badr, N. L., and Mostafa, M. G. M. (2017). Leader-based community detection algorithm for social networks. *WIREs Data Mining and Knowledge Discovery*, 7(6).

Hidalgo, C. A., Winger, B., Barabási, A. L., and Hausmann, R. (2007). The product space conditions the development of nations. *Science*, 317(5837):482–487.

Hilber, C. A. L. and Mense, A. (2021). Why have house prices risen so much more than rents in superstar cities? CEP Discussion Papers dp1743, Centre for Economic Performance, LSE.

Holland, P. W., Laskey, K. B., and Leinhardt, S. (1983). Stochastic blockmodels: First steps. *Social Networks*, 5(2):109–137.

Hu, Y. and Wang, W. (2023). *NAC: Network-Adjusted Covariates for Community Detection*. R package version 0.1.0.

Hu, Y. and Wang, W. (2024). Network-adjusted covariates for community detection. *Biometrika*, 111(4):1221–1240.

Hubert, L. and Arabie, P. (1985a). Comparing partitions. *Journal of Classification*, 2:193–218.

Hubert, L. and Arabie, P. (1985b). Comparing partitions. *Journal of classification*, 2(1):193–218.

Hugging Face (2024). all-MiniLM-L6-v2.

Johnes, G. and Hyclak, T. (1999). House prices and regional labor markets. *The Annals of Regional Science*, 33:33–49.

Kalugin, S. (2015). Les Miserables Character Network Data.

Kane, T. J., Riegg, S. K., and Staiger, D. O. (2006). School quality, neighborhoods, and housing prices. *American law and economics review*, 8(2):183–212.

Karlovčec, M., Lužar, B., and Mladenić, D. (2016). Core-periphery dynamics in collaboration networks: the case study of Slovenia. *Scientometrics*, 109(3):1561–1578.

Karrer, B. and Newman, M. E. J. (2011). Stochastic blockmodels and community structure in networks. *Physical Review E*, 83(1):016107.

Katz, J. S. (1994). Geographical proximity and scientific collaboration. *Scientometrics*, 31(1):31–43.

Knuth, D. E. (1993). *The Stanford GraphBase: a platform for combinatorial computing*. Association for Computing Machinery, New York, NY, USA.

Kojaku, S. and Masuda, N. (2017). Finding multiple core-periphery pairs in networks. *Physical Review E*, 96(5):052313.

Kojaku, S. and Masuda, N. (2018). Core-periphery structure requires something else in the network. *New Journal of Physics*, 20(4):043012.

Kosztyán, Z. T., Király, F., Katona, A. I., Csizmadia, T., and Fehérvölgyi, B. (2024). Analysis and prediction of the horizon 2020 r&d&i collaboration network. *Expert Systems with Applications*, 255:124417.

Kozitsin, I. V., Gubanov, A. V., Sayfulin, E. R., and Goiko, V. L. (2023). A nontrivial interplay between triadic closure, preferential, and anti-preferential attachment: New insights from online data. *Online Social Networks and Media*, 34-35:100248.

Kryvobokov, M. (2013). Hedonic price model: Defining neighbourhoods with thiessen polygons. *International Journal of Housing Markets and Analysis*, 6(1):79–97.

Kuhn and Max (2008). Building predictive models in r using the caret package. *Journal of Statistical Software*, 28(5):1–26.

Kwon, S., Kim, S., Tak, O., and Jeong, H. (2017). A study on the clustering method of row and multiplex housing in seoul using k-means clustering algorithm and hedonic model. *Journal of Intelligence and Information Systems*, 23(3):95–118.

Lancichinetti, A. and Fortunato, S. (2011). Limits of modularity maximization in community detection. *Physical Review E—Statistical, Nonlinear, and Soft Matter Physics*, 84(6):066122.

Lancichinetti, A., Fortunato, S., and Radicchi, F. (2008). Benchmark graphs for testing community detection algorithms. *Physical Review E - Statistical, Nonlinear, and Soft Matter Physics*, 78(4).

Lee, C. and Wilkinson, D. J. (2019). A review of stochastic block models and extensions for graph clustering. *Applied Network Science*, 4:122.

Lee, S. H., Cucuringu, M., and Porter, M. A. (2014). Density-based and transport-based core-periphery structures in networks. *Physical Review E - Statistical, Nonlinear, and Soft Matter Physics*, 89(3):032810.

Leger, J.-B., Barbillon, P., and Chiquet, J. (2021a). *blockmodels: Latent and Stochastic Block Model Estimation by a 'V-EM' Algorithm*. R package version 1.1.5.

Leger, J.-B., Barbillon, P., and Chiquet, J. (2021b). *blockmodels: Latent and Stochastic Block Model Estimation by a 'V-EM' Algorithm*. R package version 1.1.5.

Legramanti, S., Rigon, T., Durante, D., and Dunson, D. B. (2022). Extended stochastic block models with application to criminal networks. *The Annals of Applied Statistics*, 16(4):2369.

LeSage, J. and Pace, R. K. (2009a). *Introduction to spatial econometrics*. Chapman and Hall/CRC.

LeSage, J. P. and Pace, R. K. (2009b). Spatial econometric models. In *Handbook of applied spatial analysis: Software tools, methods and applications*, pages 355–376. Springer.

LeSage, J. P. and Pace, R. K. (2014). Interpreting spatial econometric models. In *Handbook of regional science*, pages 1535–1552. Springer.

Lip, S. Z. W. (2011). A fast algorithm for the discrete core/periphery bipartitioning broblem. *arXiv:1102.5511*.

Liu, L., Jones, B. F., Uzzi, B., and Wang, D. (2023). Data, measurement and empirical methods in the science of science. *Nature human behaviour*, 7(7):1046–1058.

Lorrain, F. and White, H. C. (1971). Structural equivalence of individuals in social networks. *The Journal of mathematical sociology*, 1(1):49–80.

Lu, D.-D. (2021). Leader-Based Community Detection Algorithm in Attributed Networks. *IEEE Access*, 9:119666–119674.

Lužar, B., Levnajić, Z., Povh, J., and Perc, M. (2014). Community Structure and the Evolution of Interdisciplinarity in Slovenia's Scientific Collaboration Network. *PLoS ONE*, 9(4):e94429.

Maruccia, Y., Solazzo, G., Del Vecchio, P., and Passiante, G. (2020). Evidence from Network Analysis application to Innovation Systems and Quintuple Helix. *Technological Forecasting and Social Change*, 161:120306.

McInnes, L., Healy, J., and Melville, J. (2018). UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction.

McLachlan, G. and Peel, D. (2005). Finite mixture models. *Finite Mixture Models*, pages 1–427.

McPherson, M., Smith-Lovin, L., and Cook, J. M. (2001). Birds of a feather: Homophily in social networks. *Annual Review of Sociology*, 27:415–444.

Menardi, G. (2016). A review on modal clustering. *International Statistical Review*, 84(3):413–433.

Menardi, G. and De Stefano, D. (2022). Density-based clustering of social networks. *Journal of the Royal Statistical Society. Series A: Statistics in Society*, 185(3):1004–1029.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word representations in vector space.

Miles, W. (2025). Is there a stationary home price-rent relationship in us housing markets? *International Journal of Business &; Economics (IJBE)*, 9(2):113–131.

Miłuch, O. and Kopczewska, K. (2024). Fresh air in the city: the impact of air pollution on the pricing of real estate. *Environmental Science and Pollution Research*, 31(5):7604–7627.

Morea, F. and De Stefano, D. (2023). Innovation patterns within a regional economy through consensus community detection on labour market network. In *Proceedings of the Statistics and Data Science Conference*, pages 1–6.

Morea, F. and De Stefano, D. (2025). A comprehensive framework for solution space exploration in community detection. *Scientific Reports*, 15(1):38148.

Morea, F., Soraci, A., and De Stefano, D. (2024). Mapping leadership and communities in EU-funded research through network analysis. *Open Research Europe*, 4:268.

Moretti, E. (2011). Local labor markets. In *Handbook of labor economics*, volume 4, pages 1237–1313. Elsevier.

Moretti, E. (2012). *The new geography of jobs*. Houghton Mifflin Harcourt.

Moretti, E. (2013). Real wage inequality. *American Economic Journal: Applied Economics*, 5(1):65–103.

Moro, M., Mayor, K., Lyons, S., and Tol, R. S. (2013). Does the housing market reflect cultural heritage? a case study of greater dublin. *Environment and Planning A*, 45(12):2884–2903.

Mosleh, M., Eckles, D., and Rand, D. G. (2025). Tendencies toward triadic closure: Field experimental evidence. *Proceedings of the National Academy of Sciences*, 122(27):e2404590122.

Mullins, N. C., Hargens, L. L., Hecht, P. K., and Kick, E. L. (1977). The Group Structure of Cocitation Clusters: A Comparative Study. *American Sociological Review*, 42(4):552.

Muscio, A., Ciffolilli, A., and Lopolito, A. (2022). Technological diversity in collaborative projects: insights into European research policy. *Journal of Economic Policy Reform*, 25(3):322–343.

Muto, S., Sugasawa, S., and Suzuki, M. (2023). Hedonic real estate price estimation with the spatiotemporal geostatistical model. *Journal of Spatial Econometrics*, 4(1):10.

Newman, M. (2010). *Networks: An Introduction*. Oxford University Press.

Newman, M. E. (2002). Assortative mixing in networks. *Physical Review Letters*, 89(20).

Newman, M. E. J. (2001). Clustering and preferential attachment in growing networks. *Phys. Rev. E*, 64:025102.

Newman, M. E. J. and Girvan, M. (2004). Finding and evaluating community structure in networks. *Physical Review E*, 69(2):026113.

Ng, T. L. J. and Murphy, T. B. (2021). Weighted stochastic block model. *Statistical Methods and Applications*, 30(5).

Nguyen, T. T., Nguyen, H. D., Chamroukhi, F., and McLachlan, G. J. (2020). Approximation by finite mixtures of continuous density functions that vanish at infinity. *Cogent Mathematics & Statistics*, 7(1):1750861.

Nicholls, S. (2019). Impacts of environmental disturbances on housing prices: A review of the hedonic pricing literature. *Journal of environmental management*, 246:1–10.

Noroozi, M., Rimal, R., and Pensky, M. (2021). Estimation and clustering in popularity adjusted block model. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 83(2):293–317.

Osland, L. and Thorsen, I. (2008). Effects on housing prices of urban attraction and labor-market accessibility. *Environment and Planning A*, 40(10):2490–2509.

Papadopoulos, F., Kitsak, M., Serrano, M. Á., Boguñá, M., and Krioukov, D. (2012). Popularity versus similarity in growing networks. *Nature 2012 489:7417*, 489:537–540.

Papadopoulos, F., Psomas, C., and Krioukov, D. (2014). Network mapping by replaying hyperbolic growth. *IEEE/ACM Transactions on Networking*, 23(1):198–211.

Park, J., Wood, I. B., Jing, E., Nematzadeh, A., Ghosh, S., Conover, M. D., and Ahn, Y. Y. (2019). Global labor flow network reveals the hierarchical organization and dynamics of geo-industrial clusters. *Nature Communications*, 10(1).

Peixoto, T. P. (2014). Hierarchical block structures and high-resolution model selection in large networks. *Phys. Rev. X*, 4:011047.

Peixoto, T. P. (2019). Bayesian stochastic blockmodeling. *Advances in Network Clustering and Blockmodeling*, pages 289–332.

Peixoto, T. P. (2022). Disentangling Homophily, Community Structure, and Triadic Closure in Networks. *Physical Review X*, 12(1):11004.

Perc, M. (2014). The Matthew effect in empirical data. *Journal of The Royal Society Interface*, 11(98).

Prebisch, R. (1949). The economic development of Latin America and its principal problems. *United Nations Department of Economic Affairs*.

Publications Office of the European Union (2025). Euroscivoc- european science vocabulary.

Reback, R. (2005). House prices and the provision of local public services: Capitalization under school choice programs. *Journal of Urban Economics*, 57(2):275–301.

Robinson, D. and Silge, J. (2025). Text Mining using 'dplyr', 'ggplot2', and Other Tidy Tools [R package tidytext version 0.4.3]. *CRAN: Contributed Packages*.

Robinson, J. S. D. (2017). *Text Mining with R*. O'Reilly Media, Inc.

Rohe, K., Chatterjee, S., and Yu, B. (2011). Spectral clustering and the high-dimensional stochastic blockmodel. *The Annals of Statistics*, 39(4).

Rombach, M. P., Porter, M. A., Fowler, J. H., and Mucha, P. J. (2012). Core-Periphery Structure in Networks. *SSRN Electronic Journal*.

Rosen, S. (1974). Hedonic prices and implicit markets: product differentiation in pure competition. *Journal of political economy*, 82(1):34–55.

Rossi, R. A. and Ahmed, N. K. (2014). Role discovery in networks. *IEEE Transactions on Knowledge and Data Engineering*, 27(4):1112–1131.

Rossi, R. A., Jin, D., Kim, S., Ahmed, N. K., Koutra, D., and Lee, J. B. (2020). On Proximity and Structural Role-based Embeddings in Networks: Misconceptions, Techniques, and Applications. *ACM Transactions on Knowledge Discovery from Data*, 14(5).

Rosvall, M., Delvenne, J., Schaub, M. T., and Lambiotte, R. (2019). Different Approaches to Community Detection. In *Advances in Network Clustering and Blockmodeling*, pages 105–119. Wiley.

Rozemberczki, B. (2019). Role2Vec: Learning role-based node embeddings. https://github.com/benedekrozemberczki/role2vec. Python package.

Ryu, D., Hong, J., and Jo, H. (2024). Capturing locational effects: application of the K-means clustering algorithm. *The Annals of Regional Science*, 73(1):265–289.

Schaff, S. and Thiel, P. (2023). FDZ Data description: Real-Estate Data for Germany (RWI-GEO-RED v9)-Advertisements on the Internet Platform lmmobilienScout24. Technical report, RWI Datenbeschreibung.

Schirripa Spagnolo, F., Borgoni, R., Carcagnì, A., Michelangeli, A., and Salvati, N. (2024). A spatial semiparametric m-quantile regression for hedonic price modelling. *AStA Advances in Statistical Analysis*, 108(1):159–183.

Schoch, D. (2024). CRAN: Package netUtils.

Scrucca, L. (2013). GA: A Package for Genetic Algorithms in R. *Journal of Statistical Software*, 53(4):1–37.

Scrucca, L., Fraley, C., Murphy, T. B., and Raftery, A. E. (2023). *Model-Based Clustering, Classification, and Density Estimation Using mclust in R*. Chapman and Hall/CRC.

Sedita, S. R., Caloffi, A., and Lazzeretti, L. (2020). The invisible college of cluster research: a bibliometric core–periphery analysis of the literature. *Industry and Innovation*, 27(5):562–584.

Sengupta, S. and Chen, Y. (2018). A block model for node popularity in networks with community structure. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 80(2):365–386.

Sheffrin, S. M. et al. (2008). *Fairness and market value property taxation*. Georgia State Univ., Andrew Young School of Policy Studies.

Sheppard, S. (2010). Measuring the impact of culture using hedonic analysis. *Center for Creative Community Development*, 28.

SiS FVG (2024). The Scientific and Innovation System of Friuli Venezia Giulia.

Smallbone, D., Kitching, J., Blackburn, R., and UKCES (2015). Anchor institutions and small firms in the UK: A review of the literature on anchor institutions and their role in developing management and leadership skills in small firms.

Stanley, N., Bonacci, T., Kwitt, R., Niethammer, M., and Mucha, P. J. (2019). Stochastic block models with multiple continuous attributes. *Applied Network Science*, 4(1):54.

Tang, W., Zhao, L., Liu, W., Liu, Y., and Yan, B. (2019). Recent advance on detecting core-periphery structure: a survey. *CCF Transactions on Pervasive Computing and Interaction*, 1(3):175–189.

Ter Wal, A. L. and Boschma, R. A. (2009). Applying social network analysis in economic geography: framing some key analytic issues. *The annals of regional science*, 43(3):739–756.

Tudisco, F. and Higham, D. J. (2019). A fast and robust kernel optimization method for core–periphery detection in directed and weighted graphs. *Applied Network Science*, 4(1).

Ureña-Carrión, J., Karimi, F., Íñiguez, G., and Kivelä, M. (2023). Assortative and preferential attachment lead to core-periphery networks. *Physical Review Research*, 5(4):043287.

van der Maaten, L. and Hinton, G. (2008). Visualizing high-dimensional data using t-sne. *Journal of Machine Learning Research*, 9(nov):2579–2605. Pagination: 27.

von Luxburg, U. (2007). A tutorial on spectral clustering. *Statistics and Computing*, 17(4):395–416.

Wedell, E., Park, M., Korobskiy, D., Warnow, T., and Chacko, G. (2022). Center–periphery structure in research communities. *Quantitative Science Studies*, 3(1):289–314.

Wei, C., Fu, M., Wang, L., Yang, H., Tang, F., and Xiong, Y. (2022). The research development of hedonic price model-based real estate appraisal in the era of big data. *Land*, 11(3):334.

Wheaton, W. C. and Lewis, M. J. (2002). Urban wages and labor market agglomeration. *Journal of Urban Economics*, 51(3):542–562.

White, D. R. and Reitz, K. P. (1983). Graph and semigroup homomorphisms on networks of relations. *Social Networks*, 5(2):193–234.

White, H. C., Boorman, S. A., and Breiger, R. L. (1976). Social structure from multiple networks. I. Blockmodels of roles and positions. *American Journal of Sociology*, 81(4):730–780.

Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York.

Wuchty, S., Jones, B. F., and Uzzi, B. (2007). The increasing dominance of teams in production of knowledge. *Science*, 316(5827):1036–1039.

Yakoubi, Z. and Kanawati, R. (2014). LICOD: A Leader-driven algorithm for community detection in complex networks. *Vietnam Journal of Computer Science*, 1(4):241–256.

Yanchenko, E. and Sengupta, S. (2023). Core-periphery structure in networks: A statistical exposition. *Statistics Surveys*, 17(none).

Zelnio, R. (2012). Identifying the global core-periphery structure of science. *Scientometrics*, 91(2):601–615.

Zhang, L. and Peixoto, T. P. (2020). Statistical inference of assortative community structures. *Physical Review Research*, 2(4):043271.

Zhang, X., Martin, T., and Newman, M. E. J. (2015). Identification of core-periphery structure in networks. *Physical Review E*, 91(3):032803.

Zhang, Y., Pan, R., Wang, H., and Su, H. (2023). Community detection in attributed collaboration network for statisticians. *Stat*, 12(1):e507.

Zhou, S. and Mondragón, R. J. (2004). The rich-club phenomenon in the internet topology. *IEEE Communications Letters*, 8(3):180–182.

UNIVERSITÀ
DEGLI STUDI
DI TRIESTE

La borsa di dottorato è cofinanziata con risorse dell'Unione europea, NextGeneration EU - Piano Nazionale di Ripresa e Resilienza, Missione 4 – Componente 1 – Investimento 4.1 CUP J92B22000900007