Department of Business and Management
Management and Computer Science
Chair of Algorithms

STRUCTURAL PATTERNS IN ACADEMIC COLLABORATION: DISSECTING THE DYNAMICS OF ITALIAN CO-AUTHORSHIP NETWORK

SUPERVISOR
Professor Irene Finocchi

CANDIDATE
Andrea Magri
270401

Academic Year 2023/2024

Table of Contents

Abstract

This thesis investigates the structural patterns in academic collaboration within the Italian co-authorship network, leveraging advanced network analysis techniques to understand how researchers collaborate and the implications of such collaborations. This paper studies the academic system in Italy by providing the analysis of several metrics and structural properties, such as degree distribution, clustering coefficient, and community detection. The study investigates the gender and role-based collaboration patterns, which show that there are significant differences in terms of opportunities for collaboration between male and female researchers as well as among different academic positions. The collaborative behavior varies across these categories.

Additionally, a detailed geographic analysis is applied to learn how spatial factors influence academic productivity and collaboration, highlighting the roles of major academic hubs such as Rome, Milan, and Bologna. Degree centrality, productivity, number of citations, and h-index are used in this paper to develop academic rankings for Computer Science and Economics. These rankings offer a glimpse into the relationship involving an individual's collaborative network and their academic impact.

A core aspect of this research is the investigation of whether co-authorship leads to higher academic productivity. Through propensity score matching (PSM) and regression analysis, confounding variables are controlled to isolate the effect of collaboration on productivity. The findings suggest that while co-authorship can be beneficial, excessive co-authorship rates tend to negatively impact individual academic productivity, as excessive collaborations may lead to congestion externalities that dilute individual contributions. However, it is also noted that higher h-index and citation counts, which are positively correlated with productivity, underscore the importance of maintaining a balance between co-authorship and individual research efforts to optimize academic productivity. This comprehensive analysis provides robust evidence that while some level of co-authorship enhances scholarly impact, there is a threshold beyond which it does not necessarily contribute to greater productivity and may, in fact, hinder it.

The findings highlight the need to promote more inclusive and diverse workspaces in academia. Ways to increase academic productivity and innovation include gender-equity policies, incentives for actions such as advancing early-career researchers, and new

funding models that encourage cross-institutional collaboration. Addressing these disparities and leveraging the potential of collaborative networks could stimulate diversity of thought and productivity within research networks. This thesis expands the literature on co-authorship networks by providing an overall analysis of the Italian academic environment. This analysis can be acted upon by researchers, institutions, and policymakers in their efforts to optimize research collaboration and productivity. The study illustrates how collaboration sparks academic success and provides a foundation for future research on the dynamics of academic networks.

Chapter I – Introduction

I.1- Background

Academic collaboration is a cornerstone of scientific research, enabling the sharing of knowledge, resources, and expertise among researchers. Collaboration today is commonly expressed through the co-authorship of publications, where multiple researchers share credit for a piece. Co-authorship networks, illustrating this type of collaboration, are useful tools for the structural and dynamic analysis of academic collaboration.

A co-authorship network is a social network in which nodes are authors and edges refer to papers they authored. Such networks are useful for studying the social structure of researchers, knowledge transfer within and between research communities, and collaborative endeavors on scientific productivity and innovation. This enables researchers to explore behaviors of collaboration, identify central contributors and high-impact partners, and understand the factors driving successful collaborations.

Co-authorship networks usually exhibit scale-free and small-world structures. A scale-free network is one where the degree of nodes follows a power law, with most nodes having few connections while some nodes, or "hubs," have many more connections. This suggests the existence of influential researchers central to communication and collaboration. In contrast, small-world networks exhibit both a low average path length and a high clustering coefficient, indicating that most authors can be reached from any other author through short chains of collaborations, and that authors tend to work in tightly knit groups.

The structural properties of co-authorship networks are important for several reasons. They provide insights into the effectiveness and resilience of scientific collaboration. Effective networks allow ideas and knowledge to spread quickly, whereas strong networks are invulnerable against the departure of key players. Analyzing these networks

helps identify central and influential researchers, who may be instigators of future collaborative research and innovation. Co-authorship networks can reveal imbalances in opportunities for collaboration, such as those depending on gender, academic position, or geographical location.

Gender gaps in academic collaboration are a significant concern. Research shows that compared to male researchers, female researchers typically have smaller and less visible networks, which can influence their career opportunities and access to collaborations. The position within academic ranks also affects the type and extent of collaborations, with senior researchers likely to have more extensive networks and play a key role in mentoring.

Geographical factors also play a crucial role in shaping academic collaboration. Researchers naturally find it easiest to collaborate with those in the same building or city due to logistical convenience, shared institutional affiliations, and regional research priorities. However, advancements in communication technologies are increasing remote and international collaborations, facilitating a global exchange of ideas.

Mapping the structure of the Italian co-authorship network is significant due to Italy's long tradition of scholarly activities and world-ranked research institutions. This analysis can provide insights into how historical, cultural, and institutional factors guide collaboration in the Italian network. It also elucidates regional patterns of academic productivity, highlighting major academic hubs such as Rome, Milan, and Bologna.
Furthermore, investigating the impact of co-authorship on academic productivity is a core aspect of this study. By employing advanced network analysis techniques and statistical methods, this research aims to determine whether extensive collaboration leads to higher academic output and impact. The findings from this study can inform policies and practices to enhance research productivity, foster inclusivity, and address systemic disparities in the academic landscape.

In summary, the background of this study emphasizes the importance of co-authorship networks in understanding the dynamics of academic collaboration. By analyzing these

networks, this research seeks to uncover patterns and disparities in collaboration, identify key contributors, and assess the impact of collaboration on academic productivity. The insights gained from this study can contribute to optimizing research collaboration and promoting a more equitable and productive academic ecosystem.

I.2 – Research objectives

This study aims to provide a comprehensive analysis of the Italian co-authorship network to uncover the underlying structural properties and dynamics of academic collaboration. One of the main aims is the analysis of the structural properties in the network, namely, the degree distribution and finding the connected components. The metrics represent the general structure and connectedness of the network and reveal ways that the collaborative impacts between researchers are surfed among.

Additionally, the study seeks to examine gender and role-based collaboration patterns within the network. By investigating disparities in collaboration based on gender and academic positions, the research aims to understand how these factors influence the formation and dynamics of collaborative groups. This analysis will highlight any existing inequalities and provide insights into the barriers that certain groups might face in accessing collaborative opportunities.

Another key objective is to explore the geographic dimensions of academic collaboration. The study will analyze city-level productivity and inter-city collaborations to understand how spatial factors influence collaborative behavior and academic output. This geographic analysis will identify major academic hubs and examine the role of proximity in facilitating or hindering collaboration.

The research also aims to assess the impact of collaboration on academic productivity. By employing propensity score matching (PSM) and regression analysis, the study will control for confounding variables to isolate the effect of collaboration on productivity.

This will help determine whether co-authorship positively correlates with higher academic output and impact, as measured by metrics such as h-index and citation counts.

Overall, this study seeks to provide actionable insights for researchers, institutions, and policymakers. By highlighting the role of influential researchers and institutions in fostering collaboration, uncovering gender and positional disparities, and offering a geographic perspective on academic productivity, the research aims to inform policies and practices that enhance collaboration and address disparities in the academic landscape.

I.3 – Significance of the Study

Understanding the dynamics of academic collaboration can inform policies and practices to enhance research productivity and inclusivity. This study contributes to the literature on co-authorship networks and provides actionable insights for researchers, institutions, and policymakers. By identifying key structural properties and collaboration patterns, the study aims to:

- Highlight the role of influential researchers and institutions in fostering collaboration.
- Uncover gender and positional disparities in collaborative opportunities.
- Provide a geographic perspective on academic productivity and collaboration.
- Offer recommendations for enhancing collaboration and addressing disparities.

The significance of this study lies in its potential to influence policy and practice within academic institutions. By understanding the structural properties of co-authorship networks and identifying key contributors, institutions can develop strategies to support and leverage the strengths of these networks. Furthermore, by addressing gender disparities and promoting inclusive collaboration practices, institutions can create a more equitable research environment. Finally, the geographic analysis provides valuable insights into regional disparities in academic productivity, informing targeted interventions to support collaboration in underrepresented areas.

Chapter II – Literature Review

II.1 – Co-authorship Network

Co-authorship networks are a specific type of social network where nodes represent authors, and edges represent co-authored publications. These networks provide a valuable framework for studying academic collaboration, as they map the relationships between researchers through their collaborative efforts. Co-authorship networks help reveal patterns such as the distribution of collaborative efforts, the emergence of key researchers, and the formation of research communities.

Research on co-authorship networks has shown that these networks often exhibit small-world properties, where most nodes can be reached from any other by a small number of steps. This characteristic suggests a high level of interconnectedness within the academic community, facilitating the rapid dissemination of knowledge. Additionally, co-authorship networks tend to be scale-free, meaning that a few nodes (authors) have a very high number of connections (collaborations), while most have relatively few. This scale-free property indicates the presence of influential researchers who play central roles in the network.

The analysis of co-authorship networks allows for the identification of various structural properties that can impact the dynamics of academic collaboration. These properties include degree distribution, clustering coefficient, and the identification of connected components. Understanding these properties helps researchers and policymakers identify central and influential authors, collaborative clusters, and the overall connectivity of the network.

II.2 – Social Network Analysis (SNA)

Social Network Analysis (SNA) involves the use of network theory to analyze social structures. It provides a set of tools and techniques for examining the relationships and interactions between individuals within a network. In the context of co-authorship networks, SNA helps identify key metrics that describe the network's structure and the roles of individual authors within it.

Degree Centrality: Degree centrality measures the number of direct connections a node has. In co-authorship networks, it indicates how many co-authors an author has. Authors with high degree centrality are often considered central players in the network, as they collaborate with many others.
For an undirected graph, the degree centrality *CD(i)* of a node *i* is given by the formula:

$$CD(i) = \frac{deg(i)}{n-1}$$

where *deg(i)* is the degree of the node *i* (i.e., the number of edges connected to *i*), and n is the total number of nodes in the network minus one. This normalizes the degree centrality to vary between 0 and 1, making it easier to compare across different network sizes.

Betweenness Centrality: Betweenness centrality measures the extent to which a node lies on the shortest paths between other nodes. It identifies authors who act as bridges within the network, facilitating the flow of information between different parts of the network. Authors with high betweenness centrality can influence the spread of knowledge and innovation.

$$C_b(i) = \sum_{s \neq v \neq t} \frac{\sigma_{st}(i)}{\sigma_{st}}$$

where σ_{st} is the total number of shortest paths from node *s* to node *t* and $\sigma_{st}(i)$ is the number of those paths that pass through *i*.

Closeness Centrality: Closeness centrality measures how close a node is to all other nodes in the network. It reflects the efficiency of information spread from that node. Authors with high closeness centrality can quickly access and disseminate information throughout the network.

The closeness centrality $C_C(i)$ of a node i is:

$$C_C(i) = \frac{(n-1)}{\sum_{i \neq u} d(i,u)}$$

where $d(i,u)$ is the shortest-path distance between *i* and *u*, and $n-1$ is the number of other nodes in the network.

Clustering Coefficient: The clustering coefficient measures the likelihood that an author's co-authors are also co-authors with each other, reflecting the presence of tightly knit collaborative groups. High clustering coefficients indicate the presence of research teams or close-knit academic communities.

For a node *i*, the clustering coefficient *C*(*i*) is given by:

$$C(i) = \frac{2e_i}{k_i(k_i-1)}$$

where e_i is the number of edges between the neighbors of node *i* and k_i is the degree of *i*. For undirected graphs, this ratio counts the number of edges between the neighbors of over the total possible number of edges between them.

Connected Components: Represent sub-networks where any two nodes are connected either directly or indirectly. The largest connected component typically represents the main body of the network, while smaller components indicate isolated groups or niche research areas. Identifying connected components helps understand the overall connectivity of the network and the presence of isolated sub-networks.

Community Detection Algorithms: Identify groups of nodes that are more densely connected internally than with the rest of the network. These communities often correspond to research teams, departments, or collaborative groups with shared interests. Common algorithms used for community detection include the Louvain method and the Girvan-Newman algorithm. These algorithms help identify collaborative communities and understand how researchers group together based on shared interests or institutional affiliations.

These metrics help identify influential authors, collaborative clusters, and the overall connectivity of the network. By analyzing these metrics, researchers can gain insights into the dynamics of academic collaboration and identify areas for improvement.

II.3 – Insights from Existing Literature on Co-authorship Networks

The study of co-authorship networks has revealed several key themes and patterns, including small-world properties, scale-free networks, gender disparities, geographic influence, and the impact of collaboration on productivity.

Small-world Networks: Co-authorship networks often exhibit small-world properties, characterized by short average path lengths and high clustering coefficients. This means that most researchers can be reached from any other by a small number of steps, facilitating efficient information spread and close-knit communities. Small-world networks are known for their robustness and efficiency in spreading information and ideas.

Scale-free Networks: Co-authorship networks also tend to be scale-free, characterized by a power-law degree distribution. This indicates the presence of a few highly connected nodes (hubs) and many nodes with few connections. Scale-free networks are resilient to random failures but vulnerable to targeted attacks on hubs. In the context of academic collaboration, hubs represent influential researchers who play central roles in facilitating collaboration and knowledge dissemination.

Gender Disparities: Studies have shown that female researchers often have smaller and less connected networks compared to their male counterparts. This disparity can impact their visibility, career progression, and access to collaborative opportunities. Gender-based analysis in co-authorship networks highlights the need for policies and practices that promote gender equity in academic collaboration.

Geographic Influence: Geographic proximity plays a significant role in collaboration, with researchers more likely to collaborate with others in close physical proximity. This influence is due to logistical convenience, shared institutional affiliations, and regional research priorities. However, advancements in communication technologies have facilitated remote and international collaborations, leading to a more global exchange of ideas.

Impact on Productivity: Collaboration has been linked to increased research productivity, with co-authored papers often receiving more citations than single-authored papers. The impact of collaboration on productivity is a key theme in the literature, highlighting the benefits of co-authorship in enhancing academic output and impact. Researchers with extensive collaborative networks tend to have higher productivity and influence, as measured by metrics such as h-index and citation counts.

In summary, the literature on co-authorship networks provides valuable insights into the dynamics of academic collaboration. By analyzing these networks, researchers can identify key structural properties, uncover disparities, and assess the impact of collaboration on productivity. This understanding can inform policies and practices that enhance research collaboration and promote a more equitable and productive academic environment.

Chapter III – Methodology

III.1 – Data Collection and Preprocessing

The data collection process for this study was conducted by my professors, Irene Finocchi, Alessio Martino, Fariba Ranjbar and Blerina Sinaimieri, who developed a comprehensive bibliometric network based on two distinct data sources: a list of faculty members provided by the Italian Ministry of University and Research (MUR) through a Cineca platform (referred to as the Cineca dataset) and publication information available in Semantic Scholar (referred to as the SemS dataset), as discussed in their work "Data cleaning and enrichment through data integration: networking the Italian academia", submitted to an international journal in 2024. These datasets offer complementary information that enriches the co-authorship network analysis.

III.1.1 – Data Sources

Cineca Dataset

The Cineca dataset includes data on 64,278 academics who, as of October 5, 2023, held various academic roles in Italian universities. These roles include full professors (FP), associate professors (AP), and researchers in both temporary and tenure-track assistant professor positions (RE). This dataset does not cover post-doctoral researchers or PhD students. It provides information on gender, most recent affiliation, and research area across 14 scientific fields as classified by the Ministry of University and Research (MUR). However, it does not contain bibliometric or publication data.

Table 1 below summarizes the basic statistics of the Cineca dataset.

	FP(%)	AP(%)	RE(%)	Gender (F%,M%)
Area 01 - Mathematics and informatics	27.5%	41%	31.5%	(29.6%; 70.4%)
Area 02 - Physics	24%	43.9%	32.1%	(22.9%; 77.1%)
Area 03 - Chemistry	21.5%	45.7%	32.8%	(49.7%; 50.3%)
Area 04 - Earth sciences	21%	47.6%	31.4%	(29.4%; 70.6%)
Area 05 - Biology	20.1%	43.1%	36.8%	(55.6%; 44.4%)
Area 06 - Medicine	23.5%	40.7%	35.8%	(37.7%; 62.3%)
Area 07 - Agricultural and veterinary sciences	23.9%	44%	32.1%	(42.7%; 57.3%)
Area 08 - Civil engineering and architecture	24.4%	44%	31.6%	(37.8%; 62.2%)
Area 09 - Industrial and information engineering	27.8%	35.8%	36.4%	(20.2%; 79.8%)
Area 10 - Antiquities, philology, literary studies, art history	23.5%	47.2%	29.3%	(54.3%; 45.7%)
Area 11 - History, philosophy, pedagogy and psychology	25.4%	44.9%	29.7%	(49.2%; 50.8%)
Area 12 - Law studies	35.7%	36.5%	27.8%	(39.6%; 60.4%)
Area 13 - Economics and statistics	32.6%	37.8%	29.6%	(39.3%; 60.7%)
Area 14 - Political and social sciences	23.5%	45.2%	31.3%	(41.2%; 58.8%)
Total	25.8%	41.7%	32.5%	(39.6%; 60.4%)

Table 1: Distribution of Gender and Academic Positions across Scientific Sectors in Cineca dataset [1]

Semantic Scholar Dataset:

The Semantic Scholar dataset was used to obtain publication and bibliometric information. This freely accessible dataset, obtained via an API key, includes 211,633,022 papers authored by 81,067,677 individuals. It contains extensive metadata, including authors' full and short names, aliases, citation counts, publication counts, h-indices, and affiliations. For each article, key metadata such as the list of authors, publication year, citation count, and field of study (from a list of 23 distinct fields) are available. Despite its breadth, the dataset has limitations, such as unequal representation of research fields and a preference for English-language content. Additionally, author names may be inconsistently formatted, and some important fields, like authors' affiliations and articles' fields of study, are often missing.

Since the raw data is sourced from public domains and is therefore not confidential, the thesis includes personal identifiers such as names and surnames in some sections.

[1] This table is directly sourced from "Data cleaning and enrichment through data integration: networking the Italian academia", authored by Irene Finocchi, Alessio Martino, Fariba Ranjbar and Blerina Sinaimeri. All rights reserved by the authors.

III.1.2 – Data Preprocessing and Data Linkage

Data preprocessing is essential for ensuring the accuracy and reliability of the network analysis. During the data cleaning phase, several steps were undertaken to ensure the accuracy and consistency of the dataset. One crucial step involved handling inconsistencies in the classification of academic roles.

The dataset contained various specific roles, such that temporary researcher or tenure-track assistant professor, that needed to be consolidated into three main categories: full professors (FP), associate professors (AP), and researchers (RE). This consolidation was necessary to streamline the analysis and ensure comparability across different positions. Additionally, nodes with the role 'extraordinary professor' were mapped as ‘full professor’ as part of this cleaning process, since this role is reserved for those who have achieved eligibility for the full professor category.

Furthermore, author names and institutional affiliations were standardized to reduce ambiguities and ensure consistent matching between datasets. This included handling variations in spelling, initials, and name formats (e.g., "John Smith" vs. "J. Smith").
To create a network of the Italian academic community, the Cineca dataset was used as a foundation, providing reliable information about Italian faculty members. This was supplemented with publication records from Semantic Scholar.
The data linkage process faced challenges due to data noise, as detailed below.

Entity Resolution: Matching Cineca and SemS Authors.
Initially, authors with identical full names in the Cineca dataset were eliminated, reducing the dataset to 61,371 authors from the original 64,278. Matching Cineca homonyms with SemS profiles was challenging due to frequent missing affiliation data in Semantic Scholar. The full faculty names from Cineca were then matched with the full author names in Semantic Scholar, resulting in three possible outcomes: no match, unique match, or multiple matches.

- No Match: If no corresponding match was found in Semantic Scholar, the Cineca author was discarded, affecting 3,145 names.

- Unique Match: If a unique match was found, the Cineca author was linked to the corresponding Semantic Scholar profile, which applied to 12,995 authors.

- Multiple Matches: In cases where multiple author profiles matched the same Cineca name, entities had to be disambiguated. Multiple matches could arise from true homonyms or multiple profiles for the same person. For names with up to 6 matches, the corresponding Cineca author was included; otherwise, the author was discarded. This led to the removal of 2,082 Cineca names, representing only 3% of the cleaned dataset.

To determine the exact match, information about the author's research area was used. Cineca explicitly provided the research area for each author, while Semantic Scholar inferred it from the authors' list of papers. A correspondence was manually established between the 14 Cineca research areas and the 23 Semantic Scholar fields. If there was no match, the author was discarded. If a unique match was found, the author's name was linked to the Semantic Scholar profile. In cases with multiple profiles for a Cineca name within the corresponding area, co-authors were checked for commonality to determine the best match. This process allowed matching 62% of the Cineca names, corresponding to 38,220 authors, integrating their information into the co-authorship network.

III.1.3 – Graph Construction

The co-authorship network was constructed as follows:

- Graph Representation: the co-authorship network was modeled as a graph where nodes represent individual authors, and edges represent co-authorship relationships between these authors. Each edge was weighted based on the number

of co-authored papers, allowing for a more nuanced understanding of the strength and frequency of these collaborative relationships.

- Network Attributes: each node was annotated with these attributes: To provide a comprehensive analysis, each node in the network was annotated with a variety of attributes:

  - Author's Name and ID: these identifiers ensured each node could be distinctly recognized.
  - Affiliation and City: this information helped in understanding the geographic and institutional distribution of collaborations.
  - Research Area: annotated with the specific research area of each author, categorized under broad scientific fields.
  - H-index, Citation Count, and Paper Count: these bibliometric indicators were crucial for evaluating the academic impact and productivity of each author.
  - Semantic Scholar Area: this attribute was used to cross-reference the author's main research area as identified by Semantic Scholar.
  - Position and Gender: these social attributes were critical for analyzing role-based and gender-based disparities in academic collaborations.

  Similarly, edges were annotated with:

  - Number of Co-authored Papers: indicating the volume of collaboration.
  - Citation Count of Co-authored Papers: providing a measure of the impact of the collaborative work.
  - Categories of the Papers: this helped in understanding the thematic scope of the collaborations.

- Visualization: to visually interpret the network, various visualization tools were employed. These visualizations included plotting samples of the network to highlight key nodes (authors) and clusters, providing insights into the structure and dynamics of academic collaboration. This step was crucial for identifying

visually discernible patterns and anomalies that might require further investigation.

This is how the full co-authorship network was constructed, however, to conduct a more focused analysis, the network was divided into two subgraphs: one for computer science and another for economics. By dividing the network into computer science and economics subgraphs and conducting domain-specific analyses, this study provides nuanced insights into the collaboration patterns and their implications within these two distinct research fields. The methodological rigor and comprehensive analysis ensure that the findings are robust and relevant for informing academic policies and practices.

III.2 – Tools and Libraries

1. Python

Python, a versatile and widely used programming language, served as the primary tool for data manipulation, analysis, and statistical modeling in this study. Its rich ecosystem of libraries enabled a robust analysis of complex co-authorship networks.

Below are the key Python libraries that were instrumental in this research:

- NetworkX: This library was used for creating, manipulating, and studying the structure, dynamics, and functions of complex networks. NetworkX was particularly useful for implementing various network analysis methods, such as computing degree distribution, clustering coefficients, and identifying connected components within the co-authorship network.

- Pandas: Essential for data manipulation and analysis, Pandas provided high-performance, easy-to-use data structures. It was primarily used for handling and preprocessing the large datasets involved in this study, including merging data from different sources and cleaning inconsistent entries.

- NumPy: This library supported high-level mathematical functions and multi-dimensional array operations, which were critical for handling computations involving large amounts of network data.

- Matplotlib and Seaborn: These visualization libraries were used to create a variety of plots and graphs that helped illustrate the findings of the network analysis. Matplotlib was used for customizing highly detailed plots, while Seaborn was used for generating aesthetically pleasing statistical graphics.

- SciPy: Employed for more advanced scientific computing, SciPy provided tools for formal statistical testing and engineering applications, which were crucial in the analysis phase, especially for performing regression analysis and propensity score matching.

- Statsmodels: This library was used for conducting rigorous statistical analysis, especially for estimating more sophisticated statistical models that underpinned the thesis's conclusions on co-authorship and productivity.

2. Gephi

Gephi, an open-source network visualization software, was utilized to complement the analytical capabilities of Python. Gephi is particularly renowned for its efficiency in creating large-scale network visualizations and for its user-friendly interface that facilitates the exploration of network graphs. In this study, Gephi was instrumental in:

- Visualizing Graph Networks: For both the Computer Science and Economics co-authorship networks, Gephi provided detailed visual representations that helped identify network hubs, visualize community structures, and understand the overall interconnectivity within the networks.

- Community Detection: Using algorithms like the Louvain method integrated into Gephi, this tool helped detect communities within the network, facilitating a deeper understanding of the modular structure and clustering within the academic fields studied.

Integration of Tools

The integration of Python and Gephi provided a comprehensive toolkit that supported both the quantitative and visual aspects of network analysis. Python's analytical prowess, combined with Gephi's advanced visualization capabilities, allowed for a thorough dissection of the structural patterns of academic collaboration, highlighting the pivotal roles of certain nodes, the impact of geographic locations, and the presence of gender disparities within the Italian academic landscape.

III.3 – Network Analysis Techniques

This section describes the network analysis techniques used to investigate the structure and dynamics of the Italian academic co-authorship network, in particular the computer science's subgraph and the economics' one. Various metrics and algorithms were employed to uncover insights into how researchers collaborate and the implications of these collaborations.

III.3.1 – Degree Distribution Analysis

Degree distribution analysis is fundamental to understanding the connectivity patterns within a network. By examining the degree distribution, we can identify how collaborations are spread across the network. In scale-free networks, which are typical of many social networks, a few nodes (authors) have a very high degree (many collaborators), while most nodes have a low degree. This analysis helps in identifying

influential researchers who act as central hubs within their respective subgraphs. The identification of these key players is crucial for understanding the flow of information and resources within the academic community.

III.3.2 – Clustering Coefficient Calculation

The clustering coefficient is a measure that quantifies the degree to which nodes in a graph tend to cluster together. Mathematically, for a node *i*, the clustering coefficient *C*(*i*) is given by:

$$C(i) = \frac{2e_i}{k_i(k_i - 1)}$$

where e_i is the number of edges between the neighbors of node *i* and k_i is the degree of *i*. For undirected graphs, this ratio counts the number of edges between the neighbors of over the total possible number of edges between them.

A high clustering coefficient indicates a high level of local collaboration, suggesting that researchers tend to form tightly-knit groups. This can reflect the formation of research teams or close-knit academic communities.

The average clustering coefficient of the network is calculated as:

$$C = \frac{1}{N}\sum_{i=1}^{N} C_{(i)}$$

where N is the total number of nodes in the network. This metric provides insights into the overall tendency of researchers to form clusters, which can have implications for the spread of ideas and the network's robustness against disruptions. For example, a high average clustering coefficient can indicate strong local collaboration, facilitating quick dissemination of information and resilience to node removal.

III.3.3 – Connected Components Identification

Connected components in a network represent sub-networks where any two nodes are connected directly or indirectly through other nodes. Identifying these components helps understand the overall connectivity and fragmentation of the network. In network theory, a connected component is a maximal subgraph in which any two vertices are connected to each other by paths, and which is connected to no additional vertices in the supergraph.

The largest connected component (LCC) often represents the core of the network, encompassing the majority of researchers and indicating high integration within the network. Smaller connected components might signify isolated research groups or niche areas with limited external collaboration. This analysis is crucial for identifying which parts of the network are well-integrated and which are more fragmented. It also helps in pinpointing potential areas for fostering new collaborations to enhance network connectivity.

III.3.4 – Community Detection Algorithms

Community detection algorithms are employed to identify groups of nodes that are more densely connected internally than with the rest of the network. These communities often correspond to research teams, departments, or collaborative groups with shared interests. By uncovering these communities, we can gain a deeper understanding of the network's structure and dynamics.

Louvain Method: this is a popular algorithm for community detection that optimizes modularity, a measure of the strength of division of a network into modules. The Louvain method works in two phases: modularity optimization and aggregation of nodes. The process is repeated iteratively to uncover a hierarchical structure of communities within the network.

Girvan-Newman Algorithm: this algorithm detects communities by iteratively removing edges with the highest betweenness centrality, which measures the number of shortest paths that pass through an edge. By doing so, the network breaks down into smaller, more tightly-knit communities. This method highlights key bridging nodes that connect different communities, providing insights into potential points of vulnerability in the network.

By applying these algorithms, we can identify distinct communities within the network, which can inform strategies for fostering interdisciplinary collaborations and breaking down silos within academic institutions. Understanding how researchers group together can help in designing policies to encourage more integrated and collaborative research environments.

III.4 – Gender and Role-based Analysis

This section focuses on the influence of gender and academic roles on collaboration patterns within the computer science and economics subgraphs. Understanding these patterns is vital for addressing issues of equity and inclusivity in academia.

III.4.1 – Gender-based Analysis

The gender-based analysis examines differences in collaboration patterns between male and female researchers. Metrics such as degree centrality, clustering coefficient, and betweenness centrality were analyzed by gender to identify disparities. For instance, previous studies have shown that female researchers often have smaller and less connected networks compared to their male counterparts, impacting their visibility and collaborative opportunities. By comparing these metrics within each subgraph, we can identify whether similar disparities exist in the Italian academic context and propose interventions to promote gender equity.

III.4.2 – Role-based Analysis

Role-based analysis investigates the impact of academic positions on collaboration patterns. The analysis focuses on roles such as full professors, associate professors, and researchers. Metrics were compared across different academic positions to identify disparities and patterns. For example, full professors typically exhibit higher centrality measures, acting as key nodes within the network, while researchers might have more limited collaborative networks. Understanding these dynamics is crucial for designing policies that support early-career researchers and ensure that collaboration opportunities are accessible to all academic ranks.

III.5 – Geographic Analysis

Geographic factors play a significant role in shaping academic collaborations. This section explores how spatial distribution influences collaboration patterns within the computer science and economics subgraphs.

III.5.1 – City-level Analysis

City-level analysis examines academic productivity and collaboration within major Italian cities. By aggregating data on publication and citation counts by city, we can identify major academic hubs and understand their contribution to the overall network. Cities such as Rome, Milan, and Bologna are expected to exhibit high academic productivity due to their established research infrastructure and institutional support. This analysis helps in understanding the geographic concentration of academic activities and the potential benefits of proximity for fostering collaborations.

III.5.2 – Inter and Intra-City Collaborations

Analyzing inter-city and intra-city collaborations provides insights into the regional dynamics of academic collaboration. Strong inter-city collaborations often occur between major academic hubs, reflecting their central role in the national academic network. Geographic proximity can also play a crucial role, with researchers more likely to collaborate with colleagues located in nearby cities. Mapping these collaborations helps in identifying regional clusters of academic activity and understanding how different cities contribute to the network's overall connectivity.

III.6 – Ranking Analysis

To assess the impact of co-authorship on academic productivity, rankings were developed based on four key metrics: degree centrality, productivity, citation count, and h-index. Rankings were created for two subgraphs: Computer Science and Economics. This methodology allows for a detailed comparison of individual productivity and collaborative engagement.

III.7 – Propensity Score Estimation

This section details the statistical methods used to analyze the causal impact of co-authorship on academic productivity, applied to the whole co-authorship network graph.

III.7.1 – Propensity Score Method

Propensity score estimation calculates the likelihood that an individual has a high degree centrality based on various features such as h-index, citation count, and paper count. This

method balances the treatment and control groups, ensuring a fair and unbiased comparison between high and low centrality researchers. Logistic regression models are typically used to estimate these propensity scores, which are then used to match researchers with similar characteristics but different levels of centrality. This matching process helps to control for confounding variables, isolating the effect of co-authorship on academic productivity.

III.7.2 – Matching Process

The matching process involves pairing researchers with high degree centrality (treatment group) with those who have similar propensity scores but lower centrality (control group). This method controls for confounding variables and isolates the effect of co-authorship on academic productivity. Nearest neighbor matching is a common technique used in this process, ensuring that each treated individual is matched with the closest control in terms of propensity score. This matching technique helps create a balanced dataset that mimics a randomized controlled trial, thus allowing for a more accurate estimation of the treatment effect.

III.7.3 – Regression Analysis

Ordinary Least Squares (OLS) Regression: This statistical method estimates the relationships between a dependent variable and one or more independent variables. OLS regression minimizes the sum of the squared differences between observed and predicted values. In this analysis, OLS regression was used to examine the impact of co-authorship on academic productivity. The dependent variables used are the adjusted paper count and the adjusted citation count, and the independent variables included h-index, citation count, academic position, gender, and degree centrality.

In the context of this analysis, adjusted paper count and adjusted citation count are used to account for the collaborative nature of academic publications. Adjusted paper count is calculated by dividing the total number of papers by the number of co-authors plus one

(to include the author themselves). This adjustment ensures that the productivity measure reflects the individual's contribution rather than the combined output of all co-authors. Similarly, adjusted citation count divides the total number of citations by the number of co-authors plus one.

Steps of the analysis:

Step 1: Data Preparation

The code iterates through each node in the graph and extracts relevant data, such as h-index, paper count, and citation count. It also handles cases where these values might be stored as strings representing lists.

The extracted data is stored in a list of dictionaries, which is then converted into a DataFrame.

Step 2: Propensity Score Estimation

The degree centrality for each node is calculated, and a new column (DegreeCentrality) is added to the DataFrame.

A threshold is set at the 75th percentile of the degree centrality values to define the treatment group (nodes with high degree centrality).

The logistic regression model is used to estimate propensity scores based on covariates (h-index, citation count, paper count, position, and gender). These scores indicate the likelihood of a node being in the treatment group.

Step 3: Matching

Nearest neighbor matching is performed to pair treated nodes (high degree centrality) with control nodes (low degree centrality) based on their propensity scores.

The matched pairs are combined into a new DataFrame.

Step 4: Causal Inference using Regression Analysis

The adjusted paper count and adjusted citation count are calculated by dividing the paper count and citation count by the number of co-authors plus one (to include the author themselves). This adjustment accounts for the shared productivity among co-authors.

Two OLS regression models are specified and fitted: one with adjusted paper count as the dependent variable and the other with adjusted citation count. The use of the C(Position) term ensures that "Associate Professor" is the reference category, with coefficients for "Full Professor" and "Researcher" indicating their relative effects.

This comprehensive analysis combines advanced statistical methods to rigorously evaluate the impact of co-authorship on academic productivity, providing valuable insights into the dynamics of academic collaboration.

Chapter IV – Analysis and Results

IV.1 – Introduction

This chapter delves into the analysis and results of the co-authorship network of Italian academic researchers, focusing on the subgraphs for Computer Science and Economics. The study explores various network metrics and their implications on academic collaboration, productivity, and the dynamics of research communities. For each subgraph, we focus on the giant connected component since the subgraphs are not fully connected, ensuring a meaningful and cohesive analysis.

IV.2 – Structural Properties of the Italian Co-authorship Network

IV.2.1 – Basic Information of the Subgraphs

Understanding the basic properties of each subgraph provides a foundation for further analysis. This section presents a comparative analysis of the Computer Science and Economics subgraphs within the co-authorship network. It details the structure and connectivity of each subgraph by examining metrics such as the total nodes and edges, the size of the giant connected component, and the clustering coefficient.

Computer Science Subgraph:

- Total Nodes: 1,366
- Total Edges: 2,980
- Giant Connected Component Nodes: 1,193

This indicates the largest subset of the graph in which any two nodes are connected directly or indirectly. It shows a significant proportion of the network is interconnected.

- Giant Connected Component Edges: 2,680
  The number of co-authorship relationships within the giant connected component. This indicates a high level of collaboration within the largest connected group of authors.

- Average Clustering Coefficient: 0.35
  This measures the degree to which authors in the network tend to cluster together. A value of 0.35 indicates a moderate level of clustering, suggesting that if an author collaborates with two others, those two are also likely to collaborate.

- Average Shortest Path Length: 4.41
  The average number of steps along the shortest paths for all possible pairs of network nodes. An average path length of 4.41 suggests that any author can reach another author in about 4 to 5 steps on average, reflecting a relatively close-knit network.

- Diameter: 14
  The longest of all the shortest paths between any two nodes in the giant connected component. A diameter of 14 suggests that in the worst case, an author can reach another author within 14 steps.

- Density: 0.0064
  This measures how many of the possible connections in the network are actual connections. A density of 0.0064 indicates a very sparse network, which is typical for large networks.

Economics Subgraph:

- Total Nodes: 827

- Total Edges: 1,502

- Giant Connected Component Nodes: 665
  The largest subset of the graph in which any two nodes are connected directly or indirectly. This indicates a large, interconnected group of authors within the Economics subgraph, but less connected than the Computer Science subgraph.

- Giant Connected Component Edges: 1,320
  The number of co-authorship relationships within the giant connected component. This suggests a high level of collaboration within the largest connected group of Economics authors.

- Average Clustering Coefficient: 0.21
  This measures the degree to which authors in the network tend to cluster together. A value of 0.21 indicates a lower level of clustering compared to the Computer Science subgraph, suggesting less frequent mutual collaboration among co-authors.

- Average Shortest Path Length: 6.45
  The average number of steps along the shortest paths for all possible pairs of network nodes. An average path length of 6.45 suggests that any author can reach another author in about 6 to 7 steps on average, indicating a more spread-out network compared to the Computer Science subgraph.

- Diameter: 19
  The longest of all the shortest paths between any two nodes in the giant connected component. A diameter of 19 suggests that in the worst case, an author can reach

another author within 19 steps, indicating a more dispersed network compared to Computer Science.

- Density: 0.0058
  This measures how many of the possible connections in the network are actual connections. A density of 0.0058 indicates a sparse network, which is typical for large networks.

Comparing the two subgraphs, the Computer Science network is larger and more interconnected, with a higher average clustering coefficient and a shorter average shortest path length. This suggests a tighter-knit community with more frequent collaborations among authors. The Economics network, while still interconnected, has a lower average clustering coefficient and a longer average shortest path length, indicating less frequent mutual collaborations and a more dispersed network structure.

IV.2.2 – Degree Distribution Analysis

The degree distribution plot is a critical aspect of network analysis, offering insights into the structure and dynamics of the network. Below the are the two degree distribution plots for both the Computer Science (Figure 1) and Economics (Figure 2) subgraphs, followed by a detailed examination. Similar results are also presented in "Data cleaning and enrichment through data integration: networking the Italian academia", Finocchi, Martino, Ranjbar, Sinaimeri (2024), and have been reproduced in this thesis for completeness.

Explanation of the Plot

X-Axis (degree k): This axis represents the degree of the nodes in the network. The degree of a node is the number of edges connected to it. In other words, it shows how many direct connections (or co-authorships, in this context) an author has with other authors in the network.

Y-Axis ($p(k)$): This axis represents the probability distribution of nodes with a given degree k. It indicates the frequency of nodes having exactly k connections, expressed on a logarithmic scale. This scale helps in visualizing a wide range of values more effectively, especially when dealing with networks that follow a power-law distribution.

Computer Science:

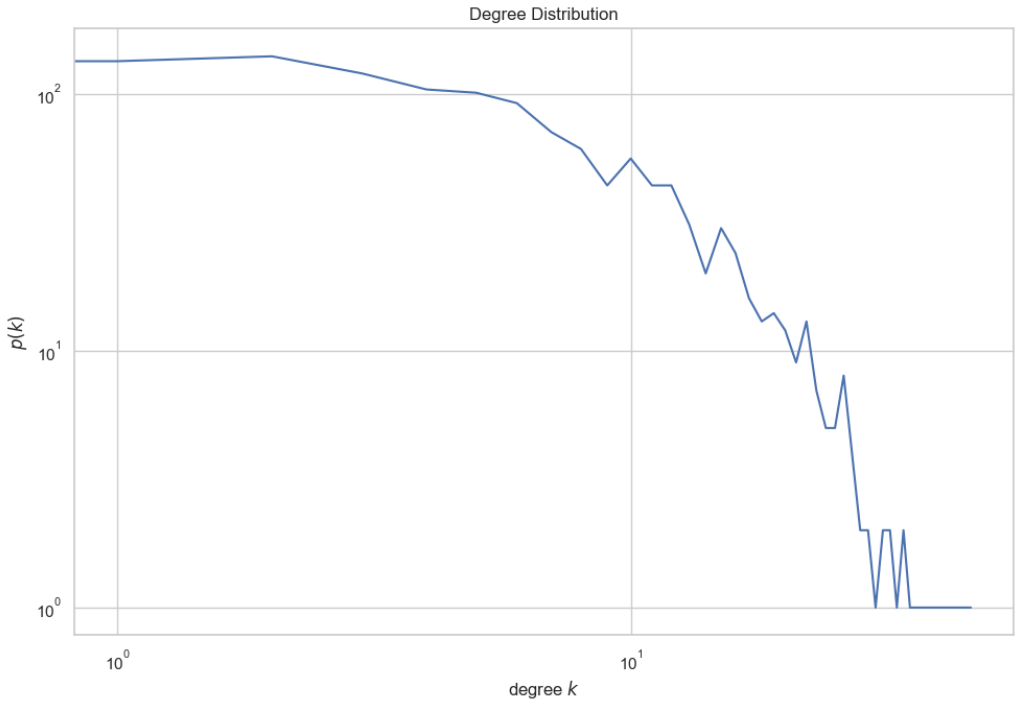

Figure 1: Degree Distribution in Computer Science Subgraph

Log-Log Scale: the plot uses a log-log scale for both axes, which is useful for identifying scale-free properties of the network, where a few nodes (authors) have a very high degree (many co-authorships), while most nodes have a low degree. In such networks, the degree distribution follows a power law, appearing as a straight line on a log-log plot.

Distribution Shape: the plot shows a downward sloping line, suggesting that the degree distribution follows a power-law distribution. This indicates the presence of hub nodes in the network — a few authors with a very high number of connections (highly collaborative authors) and many authors with fewer connections.

High-Degree Nodes: the left part of the plot represents nodes with a high degree. The steep drop-off indicates that there are very few authors with a very high number of co-authorships. These high-degree nodes are essential as they can act as bridges within the network, facilitating information flow and collaboration among disparate parts of the network.

Low-Degree Nodes: the right part of the plot represents nodes with a low degree. The flatter part of the curve at the higher degrees shows that many authors have only a few co-authorships. This is typical in academic networks where a large number of researchers contribute a few papers.

Economics:

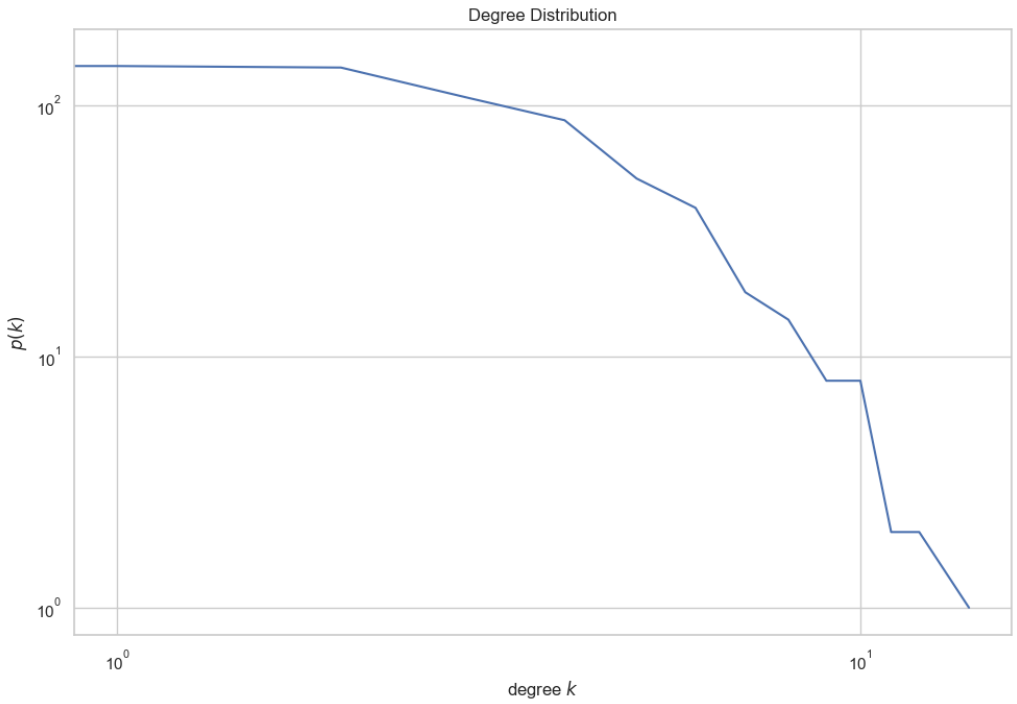

Figure 2: Degree Distribution in Economics Subgraph

Log-Log Scale: similar to the Computer Science subgraph, the plot for Economics also uses a log-log scale, highlighting the scale-free properties of the network, where a few nodes have many connections and most nodes have few.

Distribution Shape: the downward sloping line indicates a power-law distribution, suggesting the presence of hub nodes in the Economics network as well — a few authors with many connections and many authors with fewer connections.

High-Degree Nodes: the left part of the plot represents nodes with a high degree, indicating a few highly collaborative authors. These high-degree nodes are crucial for bridging the network and facilitating information flow.

Low-Degree Nodes: the right part of the plot represents nodes with a low degree, common in academic networks where many researchers have only a few co-authorships.

Implications for Network Structure

Small-World Phenomenon:
Computer Science: the presence of high-degree nodes and many low-degree nodes suggests that the Computer Science network exhibits small-world properties, meaning the network is well-connected despite its size.

Economics: similarly, the Economics network shows small-world properties due to the combination of high-degree hubs and many low-degree nodes, allowing researchers to connect through a small number of intermediaries.

Robustness and Vulnerability:
Computer Science: scale-free networks like the Computer Science subgraph are robust against random failures; removing a random node is unlikely to disrupt the network significantly. However, they are vulnerable to targeted attacks on the hubs. If key highly connected nodes are removed, the network can become fragmented quickly.
Economics: the Economics subgraph also demonstrates this characteristic. While it is robust against random failures, targeted attacks on the highly connected nodes can significantly disrupt the network.

Collaboration Dynamics

Computer Science: the degree distribution reflects the collaborative nature of the Computer Science research community. The highly connected authors are likely senior researchers or leading experts who collaborate extensively, while the numerous low-degree authors might be early-career researchers or specialists in niche areas.

Economics: the Economics subgraph shows a similar trend, with highly connected authors likely being senior researchers or leading experts, and low-degree authors being early-career researchers or specialists.

To conclude, the degree distribution analysis provides valuable insights into the collaborative structure of both the Computer Science and Economics research networks. It highlights the networks' small-world properties, robustness, and the pivotal role of highly connected authors in fostering collaboration and knowledge dissemination. Understanding these dynamics is crucial for enhancing collaborative efforts and addressing vulnerabilities within these academic communities.

IV.2.3 – Clustering Coefficient Analysis

As discussed in Chapter 3, the clustering coefficient is a key metric indicating the extent of local collaboration within the network. In the context of academic research networks, a high clustering coefficient indicates that researchers tend to form tightly-knit groups, often collaborating with each other's co-authors. This can be indicative of strong local collaboration, the formation of research teams, or the presence of closely connected academic communities.

As mentioned in section III.3.2, the clustering coefficient C_i of a node i is defined as:

$$C_i = \frac{2e_i}{k_i(k_i - 1)}$$

where e_i is the number of edges between the neighbors of node i, and k_i is the degree of node i. The average clustering coefficient for the entire network is obtained by averaging the clustering coefficients of all nodes:

$$C = \frac{1}{N}\sum_{i=1}^{N} C_i$$

where N is the total number of nodes in the network.

Interpretation for Computer Science and Economics Subgraphs

In the Computer Science subgraph, the average clustering coefficient is found to be 0.351. This relatively high value suggests that researchers in this field frequently form tight-knit collaborative groups. These clusters could facilitate the rapid exchange of ideas and innovations, fostering a vibrant academic environment. The high clustering coefficient also implies that the network is robust against random disruptions, as the removal of a few nodes is unlikely to significantly impact the overall connectivity.

In contrast, the Economics subgraph has an average clustering coefficient of 0.215. Although lower than that of Computer Science, it still indicates the presence of collaborative clusters. However, the lower value suggests that collaborations in the Economics network are less localized, possibly reflecting a more diverse array of research interests and a broader range of collaborative partnerships.

IV.2.4 – Connected Component Analysis

Connected components in a network represent sub-networks where any two nodes are connected directly or indirectly through other nodes. This analysis provides insights into the overall connectivity and potential fragmentation within the research networks.

In the Computer Science subgraph, the largest connected component (LCC) is notably extensive, comprising 1,193 nodes and 4,602 edges. This indicates a highly cohesive core network where most researchers are interconnected. Such a structure facilitates widespread collaboration and rapid dissemination of ideas across the network. The presence of a large LCC suggests that Computer Science researchers are part of a well-integrated community, enabling efficient academic exchanges and fostering collaborative opportunities.

Conversely, the Economics subgraph shows a LCC consisting of 586 nodes and 994 edges. While still substantial, this smaller component relative to the overall network size suggests a higher degree of fragmentation. The existence of several smaller connected components indicates isolated research clusters or niche areas with limited external collaboration. These fragmented components highlight opportunities for enhancing network integration by promoting interdisciplinary collaborations and connecting isolated groups to the broader research community.

Overall, understanding the structure of connected components helps identify key areas for fostering stronger connectivity and enhancing the collaborative potential within these academic fields.

IV.2.5 – Community Detection Analysis

Community detection algorithms are critical for uncovering the underlying structure of networks by identifying groups of nodes that are more densely connected internally than

with the rest of the network. These communities often correspond to research teams, departments, or collaborative groups with shared interests.

Using the Louvain method for community detection, the analysis revealed 18 distinct communities within the Computer Science subgraph and 21 communities in the Economics subgraph. The Louvain method, known for its efficiency and effectiveness, optimizes modularity to uncover hierarchical community structures. These results suggest that researchers in both fields tend to cluster into distinct groups, likely based on specific research interests, institutional affiliations, or collaborative projects. Understanding these community structures can provide valuable insights for academic institutions aiming to foster interdisciplinary research and collaboration.

The Girvan-Newman algorithm, which identifies communities by iteratively removing edges with the highest betweenness centrality, highlighted the presence of 2 primary communities in both subgraphs. This method is particularly useful for identifying key nodes that act as bridges between different communities. In both Computer Science and Economics, these bridging nodes play a crucial role in maintaining the overall cohesion of the network. They facilitate the flow of information and collaboration between otherwise disparate groups, emphasizing their importance in sustaining a connected and collaborative research environment.

Louvain Community Detection Visualization

The visualizations below depict the communities detected by the Louvain method. Each color represents a different community.

Computer Science Subgraph:

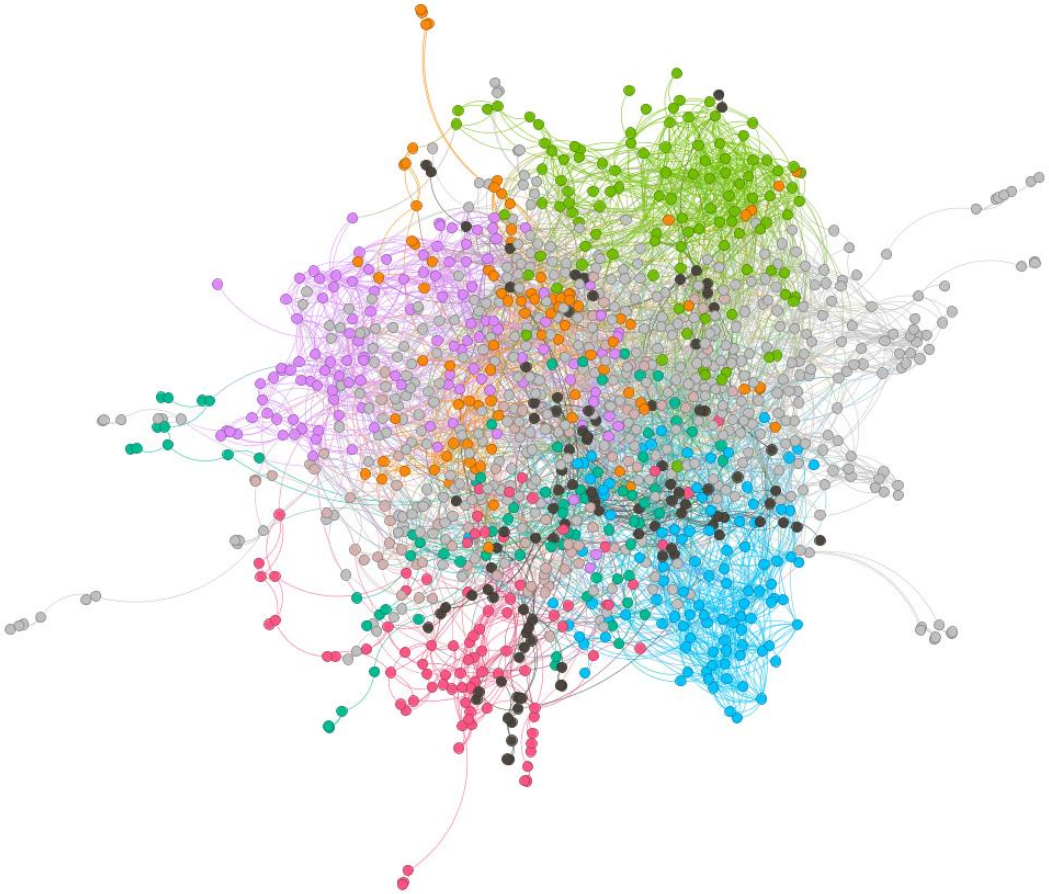

Figure 3: Community Detection of Computer Science Subgraph

Economics Subgraph:

Figure 4: Community Detection of Economics Subgraph

By analyzing these community structures, we can better understand the collaborative dynamics within each field, identify potential leaders or key influencers, and develop strategies to support and expand collaborative efforts across different research areas.

IV.3 – Gender and Role-based Collaboration Patterns

This section examines the collaboration dynamics within the academic networks of Computer Science and Economics, focusing on gender and role-based interactions to uncover disparities and trends in co-authorship.

IV.3.1 – Gender-based Collaboration Patterns

Node and Edge Distribution

The gender collaboration patterns within the Computer Science and Economics subgraphs reveal significant disparities. In the Computer Science subgraph, male researchers represent 78.37% of the nodes, indicating a substantial gender imbalance. Female researchers, accounting for only 21.63% of the nodes, are markedly underrepresented. This trend is somewhat similar in the Economics subgraph, where male researchers make up 69.80% of the nodes, while female researchers constitute 30.20%.

This gender imbalance is also reflected in the distribution of edges, which represent co-authorship connections. In the Computer Science subgraph, male-male collaborations are the most prevalent, with 2800 edges. In contrast, female-female collaborations are the least frequent, with only 314 edges. Mixed-gender collaborations, represented by 1488 edges, highlight a significant number of interactions between male and female researchers, though these do not fully mitigate the overall imbalance.

In the Economics subgraph, the pattern persists with male-male collaborations totaling 542 edges, which is significantly higher than the 92 edges representing female-female collaborations. Mixed-gender collaborations are represented by 360 edges. This disparity in collaboration patterns underscores the dominance of male researchers in academic collaborations within these fields.

Collaboration Rates

The analysis of collaboration rates provides further insights into the dynamics of gender-based collaboration. In the Computer Science subgraph, the collaboration rates are as follows:

- Male-Male: 0.0064
- Female-Female: 0.0095
- Male-Female: 0.0062

These rates suggest that female researchers exhibit a higher intra-gender collaboration rate compared to their male counterparts. This indicates that, although fewer in number, female researchers tend to form more cohesive and collaborative sub-networks within their gender group. This could reflect strong intra-gender support and collaboration among female researchers in Computer Science.

In the Economics subgraph, the collaboration rates are:

- Male-Male: 0.0065
- Female-Female: 0.0059
- Male-Female: 0.0050

These rates indicate that male researchers have a slightly higher collaboration rate compared to female researchers. The lower collaboration rate among female researchers might reflect challenges or barriers they face in forming collaborations within this field.

Gender Mixing in Collaborations

Mixed-gender collaborations are crucial for fostering an inclusive and diverse academic environment. In the Computer Science subgraph, the presence of 1488 mixed-gender edges indicates significant cross-gender collaboration. However, the overall higher male dominance suggests that female researchers may still face barriers in achieving equal participation in these collaborations.

In the Economics subgraph, the number of mixed-gender collaborations is 360, which, while significant, still highlights a disparity. The lower female collaboration rate suggests potential obstacles that limit female researchers' ability to engage in cross-gender collaborations, such as institutional biases or a lack of supportive networks.

Implications

Institutional Policies

Institutions should recognize and address the gender disparities in academic collaboration. Implementing policies that promote gender equity, such as mentorship programs, targeted funding for female researchers, and initiatives that encourage mixed-gender collaborations, can help mitigate these imbalances.

Support Networks

Creating support networks and collaborative platforms specifically for female researchers can enhance their visibility and participation in academic collaborations. Such initiatives can empower female researchers to expand their networks and increase their collaborative engagements.

Cultural Change

Beyond policies, fostering a cultural change within academic institutions that values and promotes diversity in collaboration is essential. This involves recognizing the contributions of female researchers and actively working to dismantle systemic barriers that hinder their full participation.

The gender disparities in the Italian co-authorship network highlight significant challenges and opportunities for promoting gender equity in academic collaborations. While male researchers dominate in numbers and connectivity, female researchers exhibit higher intra-gender collaboration rates in certain fields, indicating strong cohesion within their sub-networks. Addressing the underlying causes of these disparities through targeted policies and cultural shifts is crucial for creating a more inclusive and productive academic environment.

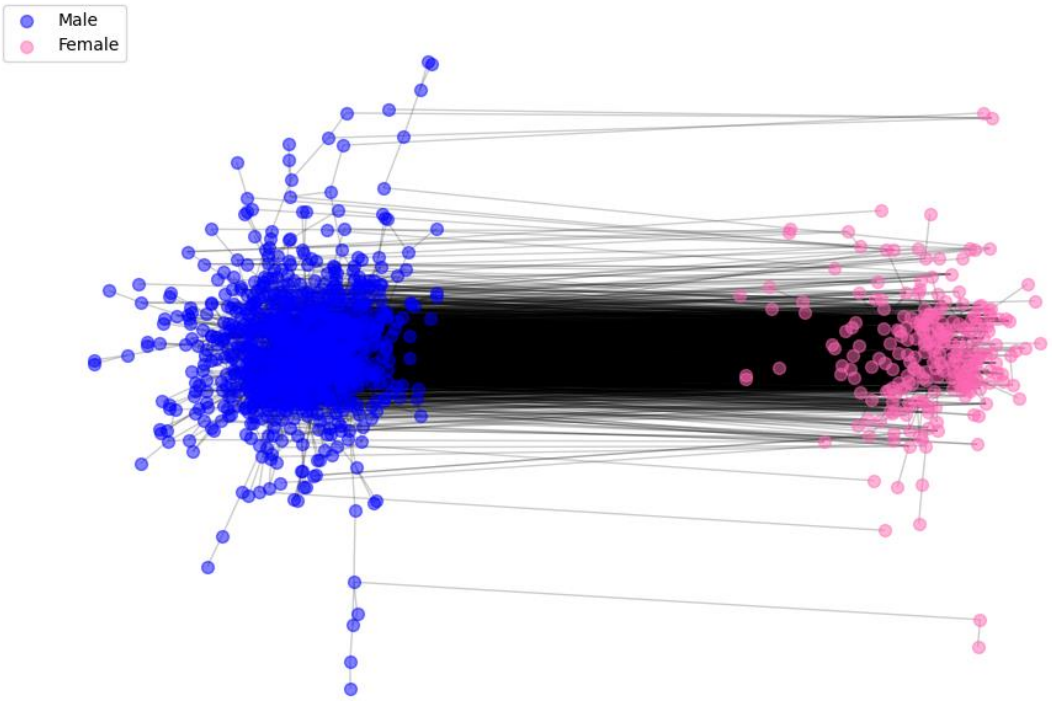

Figure 5: Gender-Based Bipartite Visualization of the Computer Science Subgraph

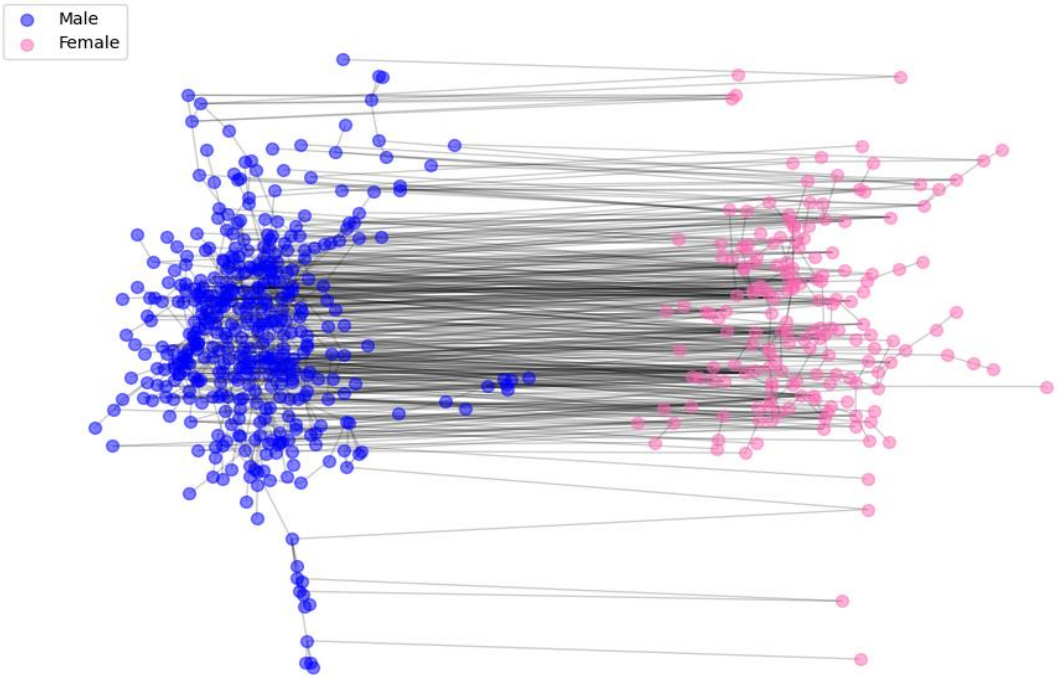

Figure 6: Gender-Based Bipartite Visualization of the Economics Subgraph

These network visualizations illustrate the co-authorship relationships within the fields of Computer Science and Economics, separated by gender.
As indicated in both subgraphs, blue nodes represent male researchers, and pink nodes represent female researchers. The density and distribution of the connections highlight the prevalent male-male collaborations while also depicting the extent of mixed-gender and female-female collaborations. The visualizations underscore the gender disparities in academic collaborations, with noticeable differences in collaboration patterns and node distribution between the two fields.

IV.3.2 – Role-based Collaboration Patterns

Understanding the dynamics of role-based collaboration in academic networks is crucial for assessing the structural integration and productivity within different fields. This analysis focuses on the role-based collaboration patterns within the Italian co-authorship networks in Computer Science and Economics. By examining the distribution of academic positions (full professor, associate professor and researcher) and their collaboration frequencies, we gain insights into the hierarchical and integrative nature of these networks.

Computer Science Faculty Distribution

- Full Professors: 233 (19.53%)
- Associate Professors: 430 (36.04%)
- Researchers: 530 (44.43%)

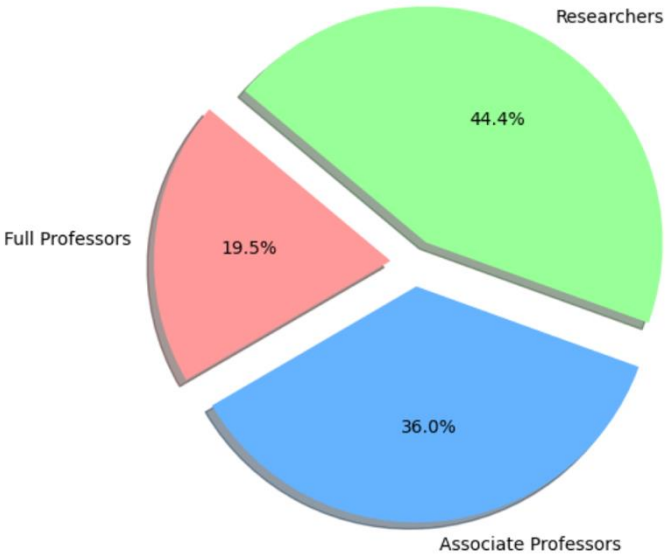

Figure 7: Distribution of academic roles in Computer Science

The Computer Science network is predominantly composed of researchers (44.43%), followed by associate professors (36.04%) and full professors (19.53%).
This distribution suggests a strong foundation of early-career researchers (researchers) and mid-career academics (associate professors), with a relatively smaller proportion of senior academics (full professors), that could be due to the relative youth of this subject.

Collaboration Patterns

	Full Professor	Associate Professor	Researcher	Total
Full Professor	506.0	593.5	327.5	1427.0
Associate Professor	593.5	821.0	496.5	1911.0
Researcher	327.5	496.5	440.0	1264.0
Total	1427.0	1911.0	1264.0	4602.0

Table 2: Collaboration matrix for the Computer Science network [2]

[2] Note on edge counting:
For edges where both nodes have the same role (diagonal cells), the count is incremented by 1. For edges connecting nodes with different roles (off-diagonal cells), the count is incremented by 0.5 for each role. This avoids double-counting, ensuring each edge is counted exactly once in the totals.

Full professor collaboration: full professors have a total of 1427 collaborations, indicating significant engagement both within their own rank and across other ranks. The high count of collaborations with associate professors (593.5) and researchers (327.5) highlights their pivotal role in bridging different academic levels.

Associate professor collaboration: associate professors exhibit the highest total number of collaborations (1911.0), reflecting their central position in the network. Their strong ties with full Professors (593.5) and researchers (496.5) suggest that they play a crucial role in facilitating cross-rank collaborations.

Researcher collaboration: researchers, while being the most numerous, show a lower total number of collaborations (1264.0) compared to associate Professors. Their collaborations are more evenly distributed between their peers and higher ranks, emphasizing their role in knowledge transfer and network integration.

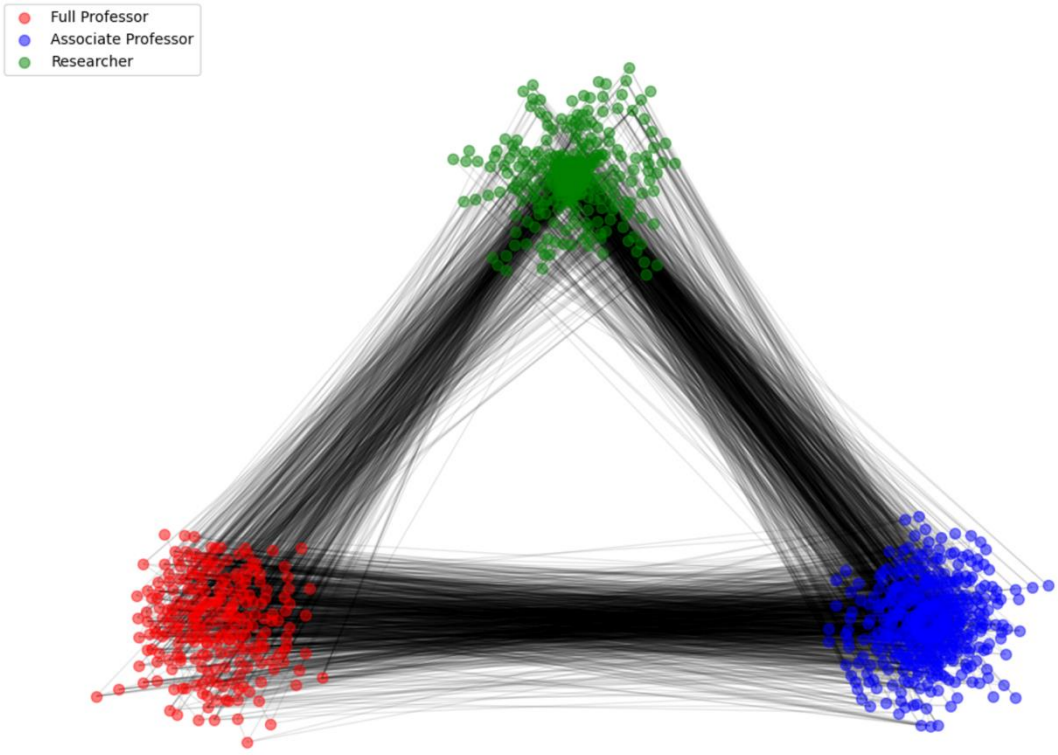

Figure 8: Role-based Tripartite Network of Computer Science

Economics Faculty Distribution

- Full Professors: 229 (39.08%)
- Associate Professors: 223 (38.05%)
- Researchers: 134 (22.87%)

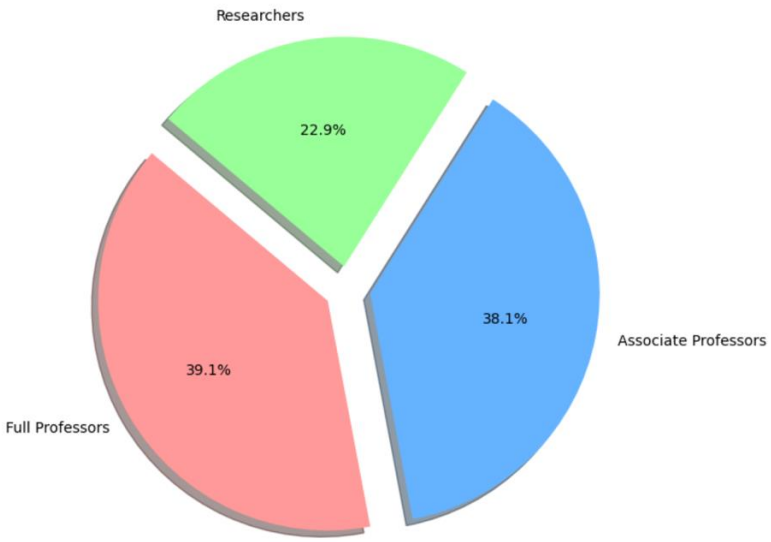

Figure 9: Distribution of academic roles in Economics

The Economics network has a more balanced distribution among full professors (39.08%), associate professors (38.05%), and researchers (22.87%). This indicates a relatively stable hierarchy with significant representation across all academic ranks.

Collaboration Patterns

	Full Professor	Associate Professor	Researcher	Total
Full Professor	261.0	153.0	69.0	483.0
Associate Professor	153.0	132.0	63.5	348.5
Researcher	69.0	63.5	30.0	162.5
Total	483.0	348.5	162.5	994.0

Table 3: Collaboration matrix for the Economics network [3]

Full professor collaboration: full professors in Economics have fewer total collaborations (483.0) compared to their Computer Science counterparts, but still demonstrate strong engagement with both associate professors (153.0) and researchers (69.0). This balanced interaction suggests a collaborative environment where senior academics actively engage with less senior colleagues.

Associate professor collaboration: associate professors have a total of 348.5 collaborations, showing robust interactions with full professors (153.0) and researchers (63.5). Their role as a collaborative bridge is evident, facilitating integration across ranks.

Researcher collaboration: researchers, despite being the least numerous, maintain active collaboration networks (162.5 total). Their interactions with associate professors (63.5) and full professors (69.0) are vital for their professional development and integration into the academic community.

[3] Note on edge counting:
For edges where both nodes have the same role (diagonal cells), the count is incremented by 1.
For edges connecting nodes with different roles (off-diagonal cells), the count is incremented by 0.5 for each role. This avoids double-counting, ensuring each edge is counted exactly once in the totals.

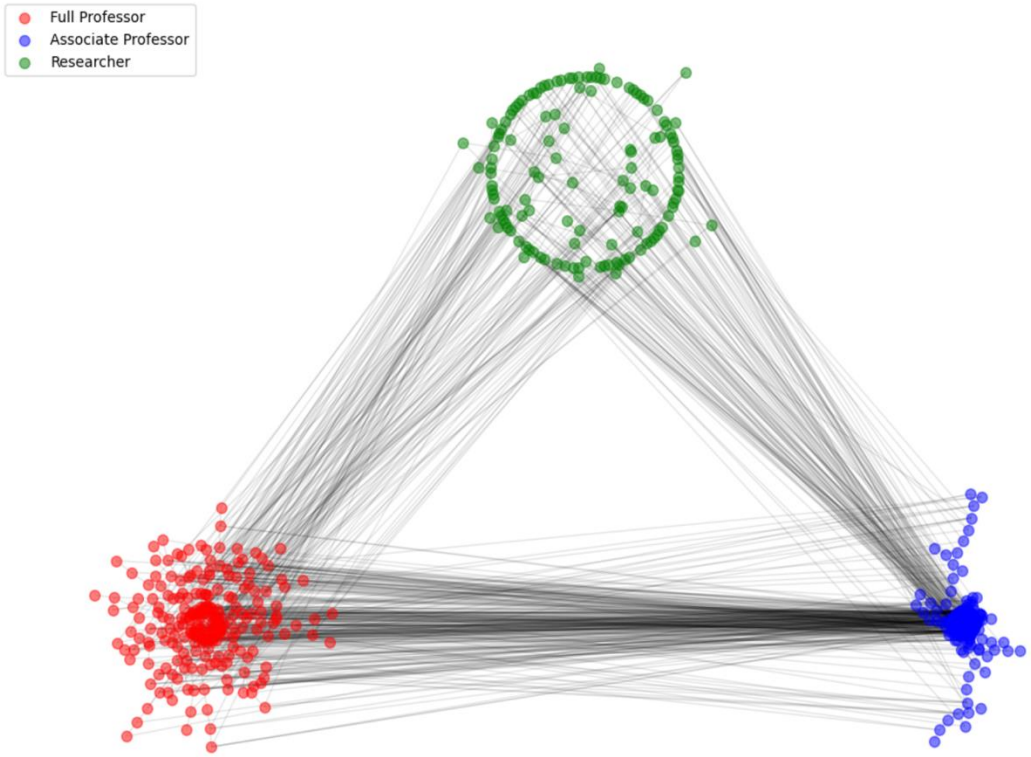

Figure 10: Role-based Tripartite Network of Economics

Comparative Insights

Network Density: The Computer Science network is denser, with a higher total number of collaborations (4602.0) compared to the Economics network (994.0). This suggests that Computer Science researchers engage more frequently in collaborative activities, potentially driven by the interdisciplinary nature of the field.

Role Integration: In both networks, Associate Professors play a central role in fostering collaborations across ranks. However, the higher proportion of Researchers in Computer Science indicates a more dynamic and expansive early-career researcher base compared to Economics.

Collaboration Balance: The Economics network shows a more balanced collaboration distribution among roles, reflecting a cohesive and integrated academic structure. In contrast, the higher collaboration counts in Computer Science suggest an active, but potentially more hierarchical, collaboration structure.

To conclude this section, the role-based analysis of the Italian co-authorship networks in Computer Science and Economics reveals distinct collaboration patterns influenced by the distribution and integration of academic positions.
In both fields, mid-career researchers (associate professors) act as key connectors, facilitating cross-rank collaborations. The higher density and expansive early-career researcher base in Computer Science suggest a more dynamic and rapidly evolving network, while the balanced role distribution in Economics indicates a stable and cohesive academic structure. These insights can inform targeted policies to enhance collaboration and support researchers at different career stages.

IV.4 – Geographic Analysis

Geographic analysis provides insights into the spatial distribution of academic productivity and collaboration within the Italian co-authorship network. By examining city-level collaboration patterns and inter-city networks, we can understand how geographic proximity and institutional affiliations influence academic collaboration.

IV.4.1 – City Level Productivity Patterns

City-level analysis examines academic productivity and collaboration within major Italian cities. By aggregating data on publication and citation counts by city, we can identify major academic hubs and understand their contribution to the overall network. Cities such as Rome, Milan, and Bologna are expected to exhibit high academic productivity due to their established research infrastructure and institutional support. This

analysis, indeed, helps in understanding the geographic concentration of academic activities and the potential benefits of proximity for fostering collaborations.

Academic productivity is analyzed at the city level, focusing on the ten cities that exhibit higher academic productivity, with substantial numbers of publications, citations, and cumulative h-index scores.

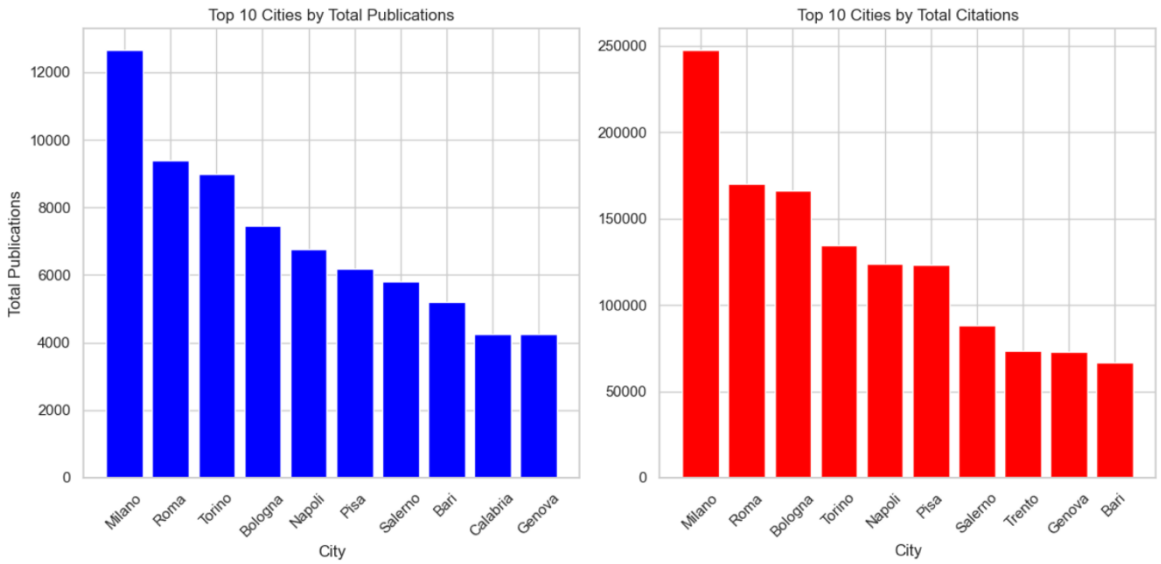

Figure 11: Top 10 Cities by Total Publications and Total Citations in Computer Science

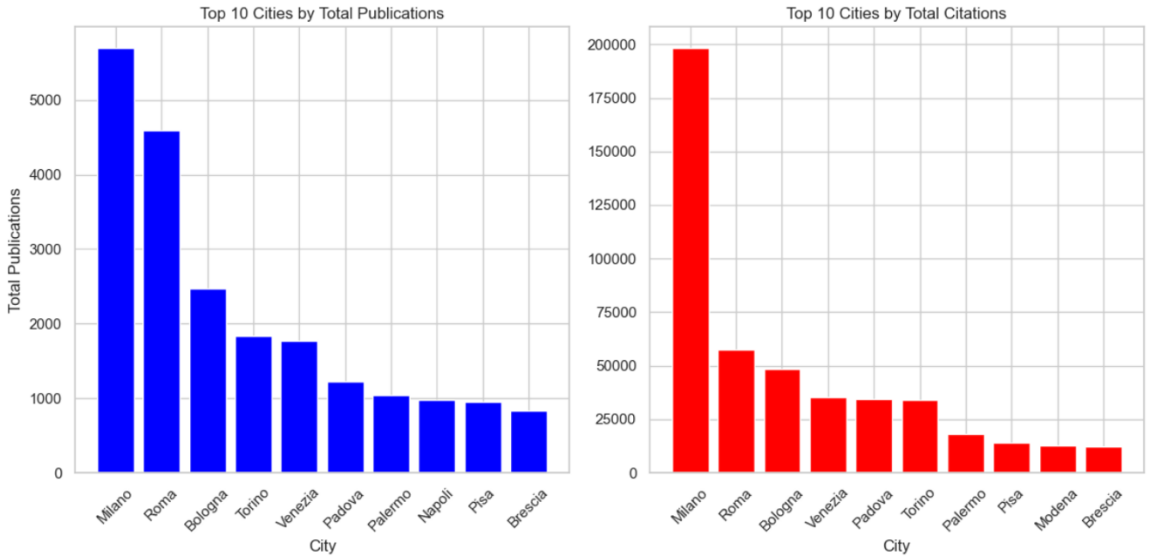

Figure 12: Top 10 Cities by Total Publications and Total Citations in Economics

Total Publications: this visualization shows the leading cities based on the number of academic papers produced, indicating the volume of research activity.

Total Citations: reflects the impact and reach of research conducted in these cities, as measured by the number of times their publications are cited.

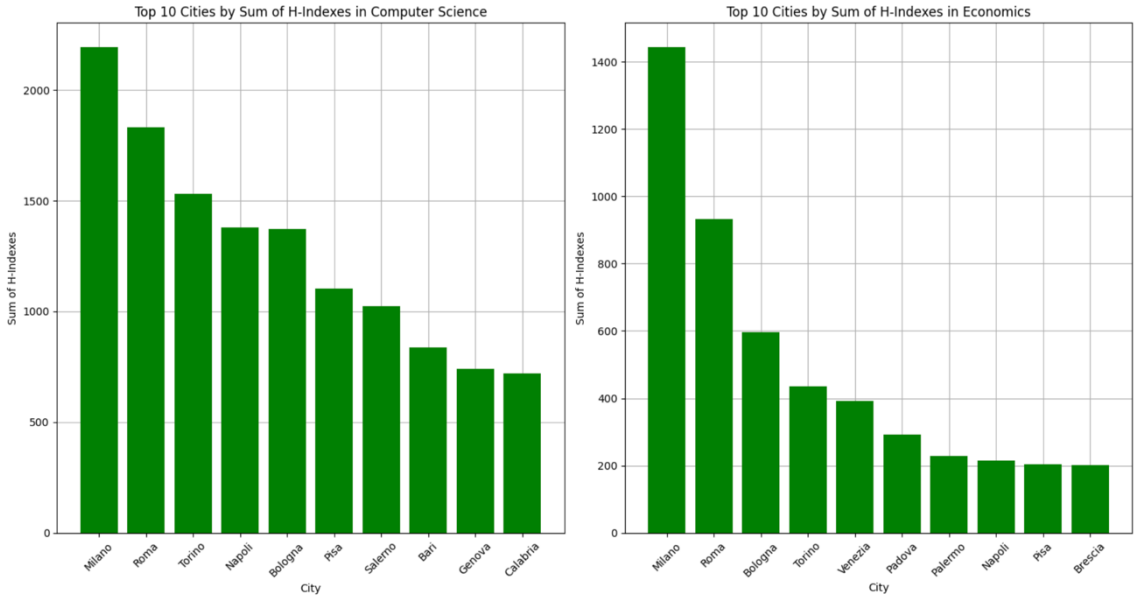

Figure 13: Top 10 Cities by Sum of H-Index in Computer Science and Economics

Sum of H-Indexes: Illustrates the overall research influence and productivity of researchers in these cities, combining both quantity and quality of publications.

Key Findings:

- Rome: as the capital city, Rome hosts numerous prestigious universities and research institutions, contributing to its high level of academic productivity. The city's strong academic infrastructure supports a significant output of publications and high citation counts. Rome’s academic ecosystem is characterized by a wide range of research activities, multidisciplinary collaborations, and significant institutional support, making it a central hub for academic research in Italy.

- Milan: known for its leading academic and research institutions, Milan also shows high levels of academic output. The city's collaborative environment is fostered by its concentration of universities and research centers, resulting in substantial publication and citation metrics. Milan benefits from a vibrant academic community and strong industry linkages that enhance research opportunities and outputs.

- Bologna: this city is another significant academic hub, with a strong tradition of research and education. Bologna's academic network is characterized by a high number of publications and notable h-index scores, reflecting the impact of its research activities. The University of Bologna, one of the oldest in the world, plays a pivotal role in driving the city’s research productivity and fostering a culture of academic excellence.

To conclude, the analysis shows that Rome and Milan consistently appear at the top across all three metrics, demonstrating their comprehensive strength in both research output and impact. These cities not only produce a high volume of publications but also generate significant citations, highlighting the influence and quality of their research. Bologna maintains a strong position, particularly in the sum of h-indexes, indicating influential research. Bologna’s academic strength is reflected in its ability to generate high-impact research, contributing substantially to the academic prestige of Italy. Other cities like Turin, Venice, and Padua also show significant academic productivity, contributing to Italy’s diverse and robust research environment. While not as dominant as Rome and Milan, these cities play crucial roles in their respective regions, fostering local academic communities and contributing to national research outputs.

IV.4.2 – Inter-City Collaboration Patterns

Inter-city collaborations are mapped to understand the regional dynamics of academic collaboration. Strong collaborative ties are observed between major academic hubs, with smaller cities often collaborating with nearby larger cities.

The analysis reveals that major academic hubs such as Rome, Milan, and Bologna exhibit strong collaborative ties with other cities, reflecting their pivotal role in the national academic network. These cities act as central nodes or hubs, facilitating extensive collaborations that extend beyond their immediate geographic boundaries. This centrality in the network underscores their importance not only as producers of a high volume of academic work but also as key facilitators of knowledge exchange and collaborative research efforts across Italy.

Rome, being the capital city, benefits from a concentration of prestigious universities and research institutions, which naturally positions it as a central hub in the academic network. Similarly, Milan, known for its leading academic and research institutions, and Bologna, with its strong tradition of research and education, also serve as major hubs. These cities not only attract significant academic talent but also foster an environment conducive to inter-city collaborations. This is evident from the strong collaborative ties they maintain with both nearby and distant cities.

Geographic proximity plays a crucial role in fostering academic collaborations. The collaboration networks show that researchers are more likely to collaborate with colleagues located in nearby cities. This is likely due to the ease of maintaining regular contact and the ability to leverage existing institutional relationships. Proximity reduces the logistical challenges associated with collaboration, such as travel time and costs, making it easier for researchers to engage in joint projects and share resources.

Visualizations:

1. City Collaboration Matrix:

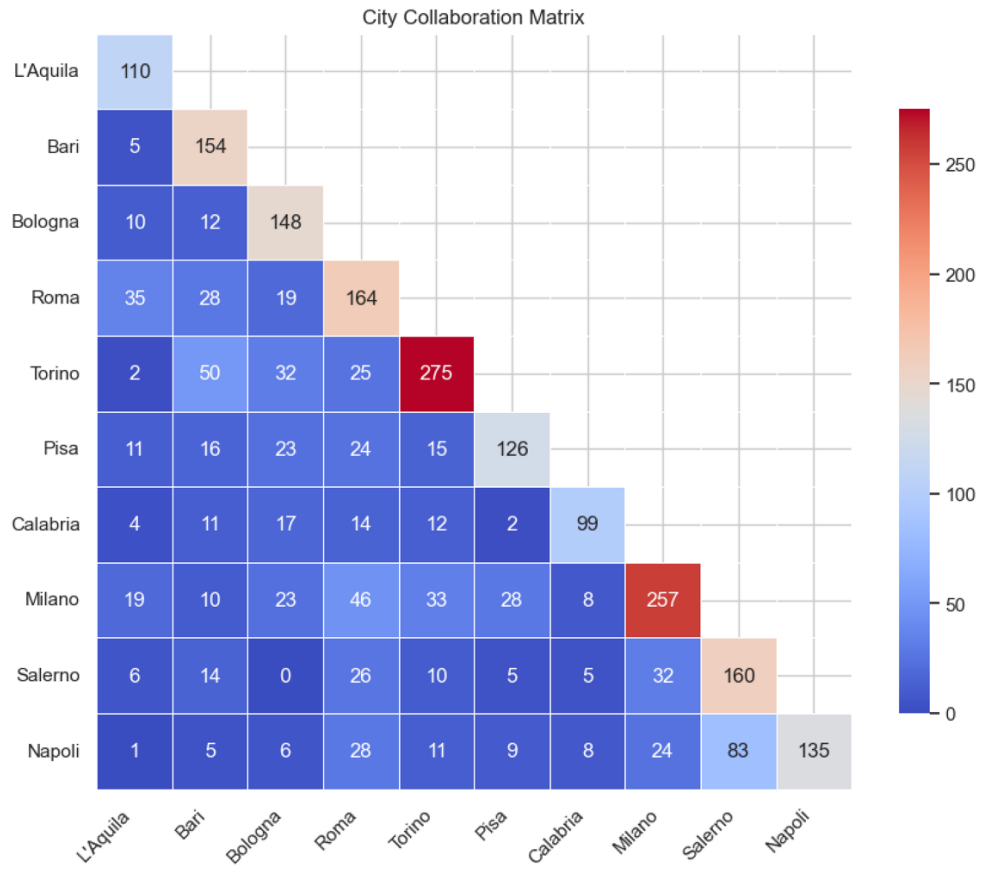

Figure 14: City Collaboration Matrix

This matrix helps to visualize these inter-city collaborations, showing the number of collaborative links between different cities. For instance, the strong linkages between Milan and other cities highlight its role as a central hub. Similarly, Rome and Bologna also display numerous collaborative ties, further emphasizing their importance in the academic network.

2. Subgraph of Top 10 Cities by Inter-City Collaborations:

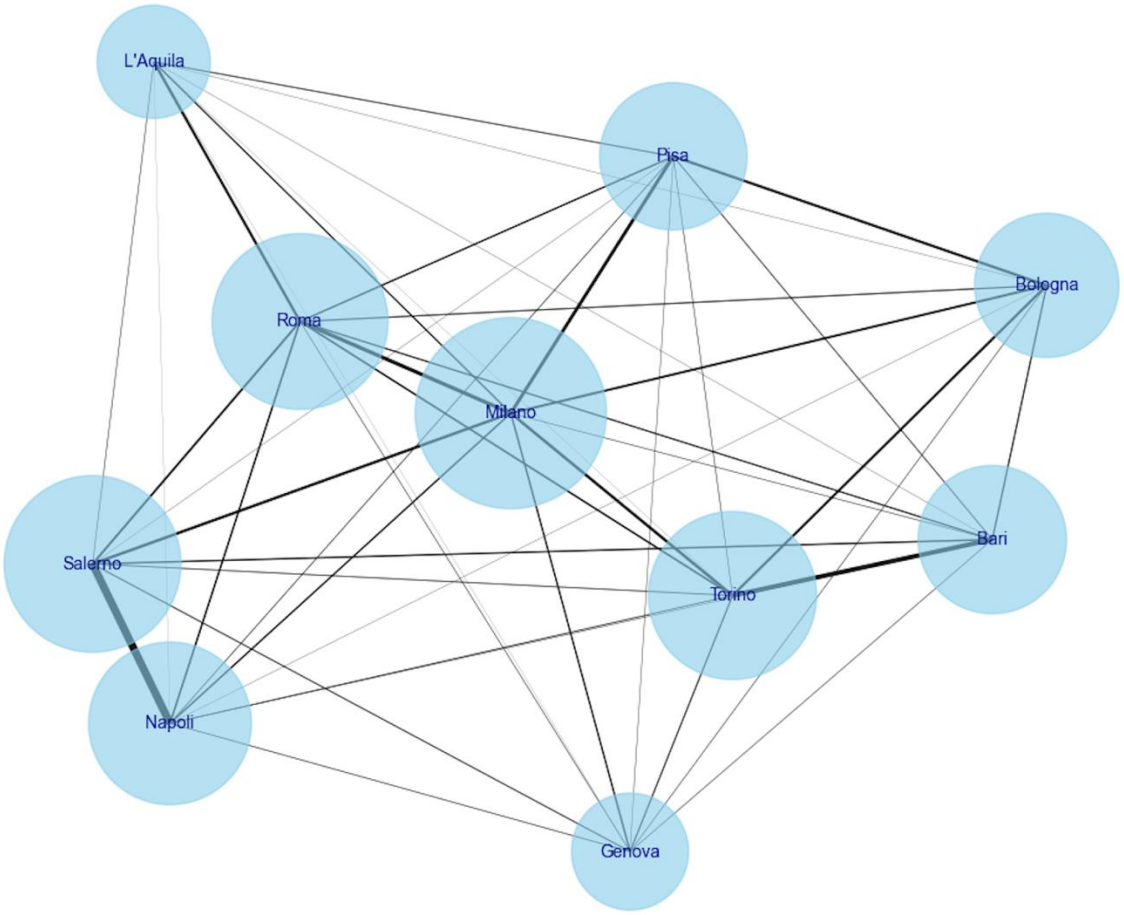

Figure 15: Subgraph of Top 10 Cities by Inter-City Collaborations

3. Geographic Collaboration Network of Top Italian Cities:

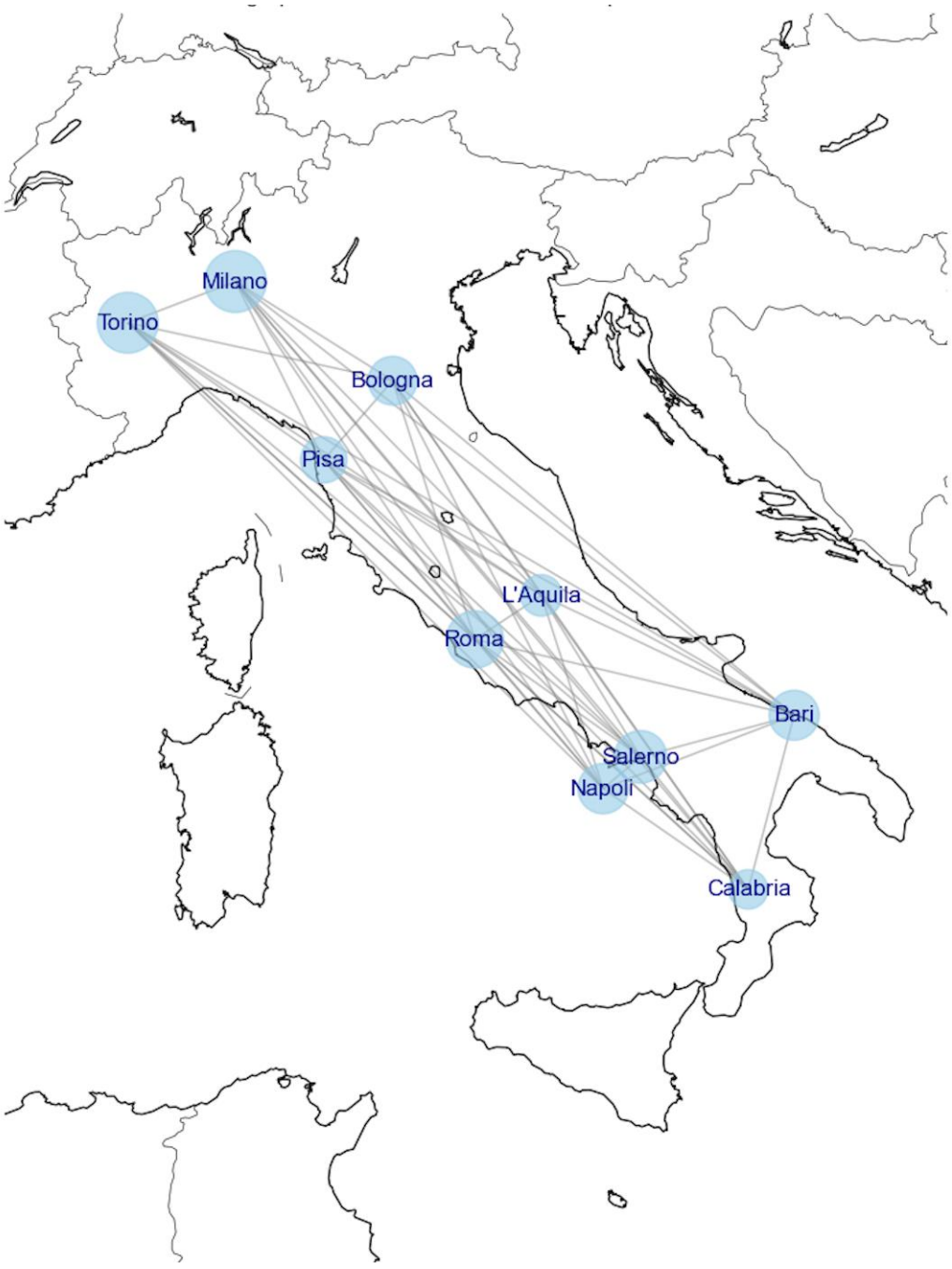

Figure 16: Geographic Collaboration Network of Top Italian Cities

The Top 10 Cities by Inter-City Collaborations (Figure 15) and the Geographic Collaboration Network of Top Italian Cities (Figure 16) provide additional insights into the structure of these collaborations. The geographic network map illustrates how cities are interconnected, with thicker lines representing stronger collaborative ties.
This visual representation helps to understand the spatial distribution of collaborations and the prominence of major academic hubs in fostering these connections.

In summary, the analysis of inter-city collaboration networks underscores the importance of major cities like Rome, Milan, and Bologna in Italy's academic landscape. Their roles as central hubs facilitate a wide range of collaborations that contribute to the strength and dynamism of the national academic network. At the same time, the significance of geographic proximity in promoting collaborations highlights the need for policies that can further enhance these networks, particularly by supporting smaller and more remote cities in forming robust collaborative links.

Analysis and Implications
Regional disparities in academic collaboration are evident from the analysis. While major cities benefit from strong academic infrastructures and high levels of collaboration, smaller and more remote cities may face challenges in forming extensive collaborative networks. Addressing these disparities requires targeted policies to support research and collaboration in less connected areas. Initiatives such as providing funding for travel, creating joint research programs, and establishing regional research hubs can help mitigate these disparities.
Institutions in major academic hubs should leverage their position to foster inter-city collaborations by providing platforms and resources to facilitate partnerships with smaller institutions. This can include hosting conferences, workshops, and collaborative research projects that involve participants from various cities. By doing so, these institutions can help create a more balanced and inclusive academic network, ensuring that the benefits of collaboration and academic exchange are more widely distributed across the country.
To conclude, the geographic analysis highlights the significant role of major cities in Italy's academic collaboration network. Cities like Rome, Milan, and Bologna serve as central hubs, fostering extensive collaborative networks that are crucial for the dynamism

and productivity of the national academic landscape. However, there is a need to address regional disparities by supporting smaller and remote institutions. Enhancing inter-city collaborations can lead to a more inclusive and productive academic environment, benefiting the entire academic community in Italy.

IV.5 – Academic Rankings

In analyzing academic productivity and collaboration within Computer Science and Economics, rankings are developed based on four key metrics: degree centrality, productivity, citation count, and h-index. These rankings are created for two subgraphs, representing the fields of Computer Science and Economics. This methodology facilitates a detailed comparison of individual productivity and collaborative engagement among top-performing researchers. By analyzing these indicators, the aim is to identify key contributors, understand their influence on the academic network, and highlight the mechanisms through which knowledge is disseminated and expanded in these critical areas of study.

Degree centrality measures the number of direct connections a node has in a network, indicating the extent of an individual's collaboration within the academic community. In Computer Science, Giovanni Semeraro leads with the highest degree centrality, reflecting his extensive collaborative network. Similarly, in Economics, Giorgio Brunello ranks highest, demonstrating his significant engagement in co-authorships.

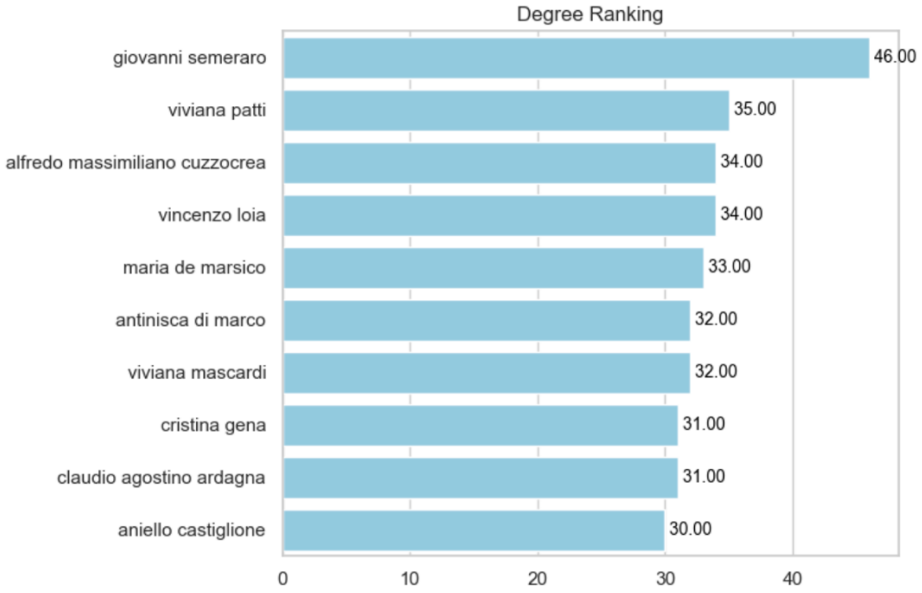

Figure 17: Degree Ranking for Computer Science

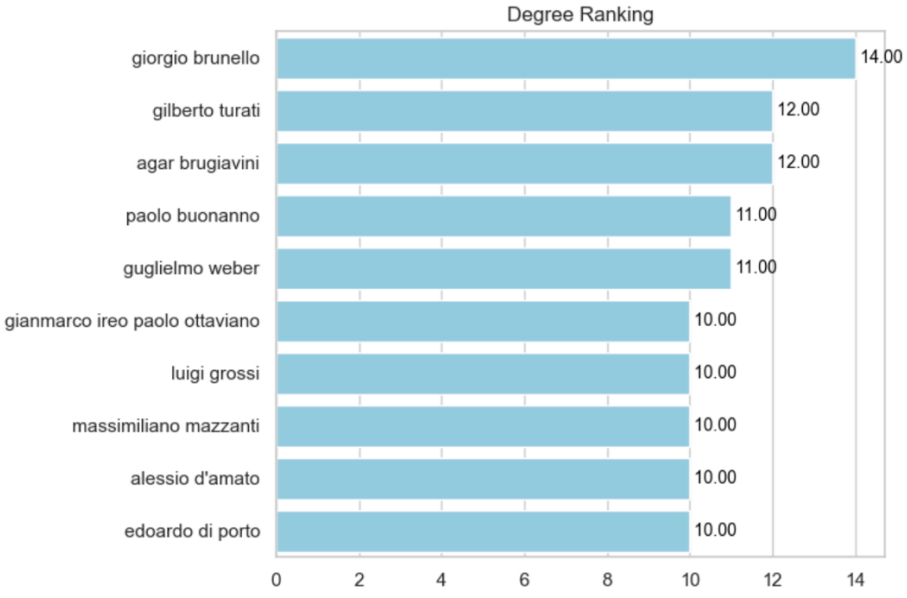

Figure 18: Degree Ranking for Economics

Productivity ranking is based on the number of publications attributed to each researcher. Giovanni Semeraro again stands out in Computer Science with the highest productivity score, showcasing his prolific contribution to research. In Economics, Massimiliano Mazzanti leads in productivity, indicating his substantial research output.

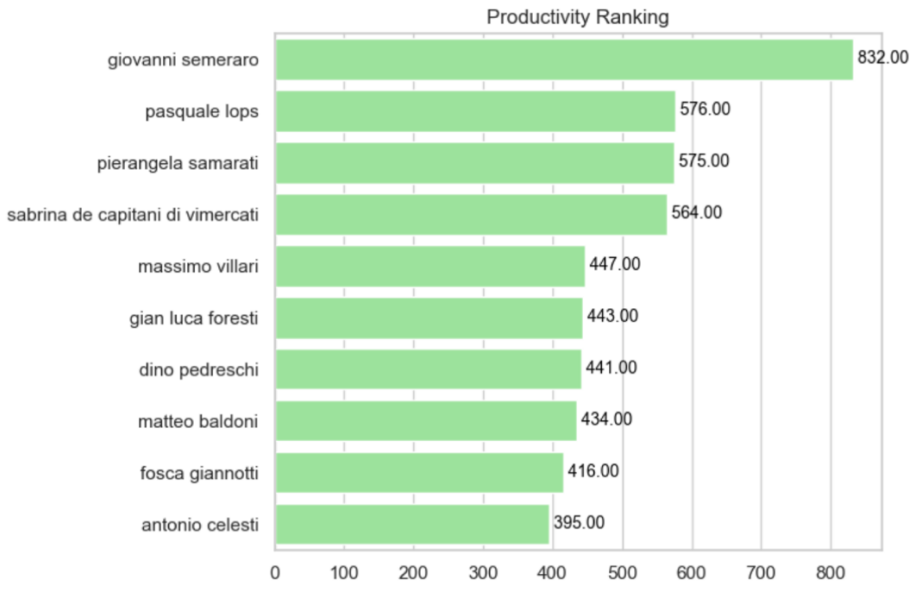

Figure 19: Productivity Ranking for Computer Science

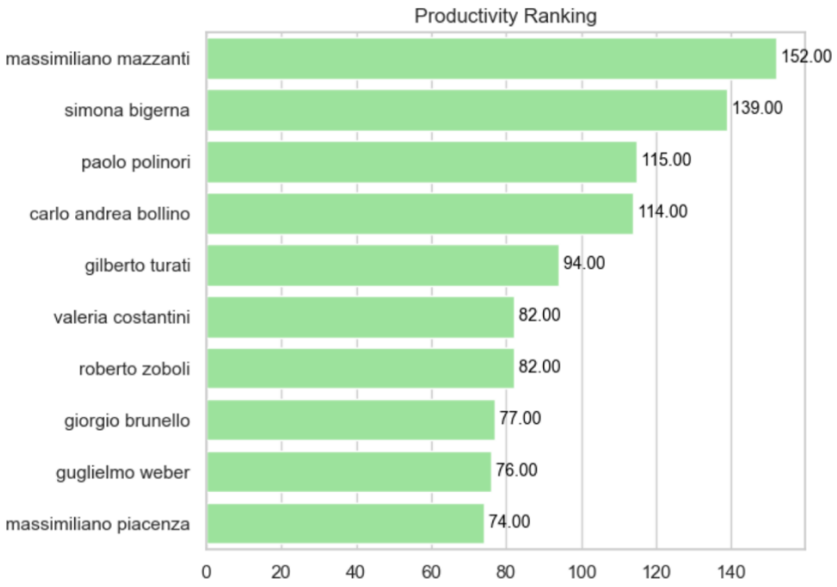

Figure 20: Productivity Ranking for Computer Science

Citation count is a measure of the impact and influence of a researcher's work, based on the number of times their publications are cited by others. Pierangela Samarati tops the citation ranking in Computer Science, highlighting the significant influence of her research. In Economics, Massimiliano Mazzanti again takes the lead, reflecting the wide recognition and impact of his work.

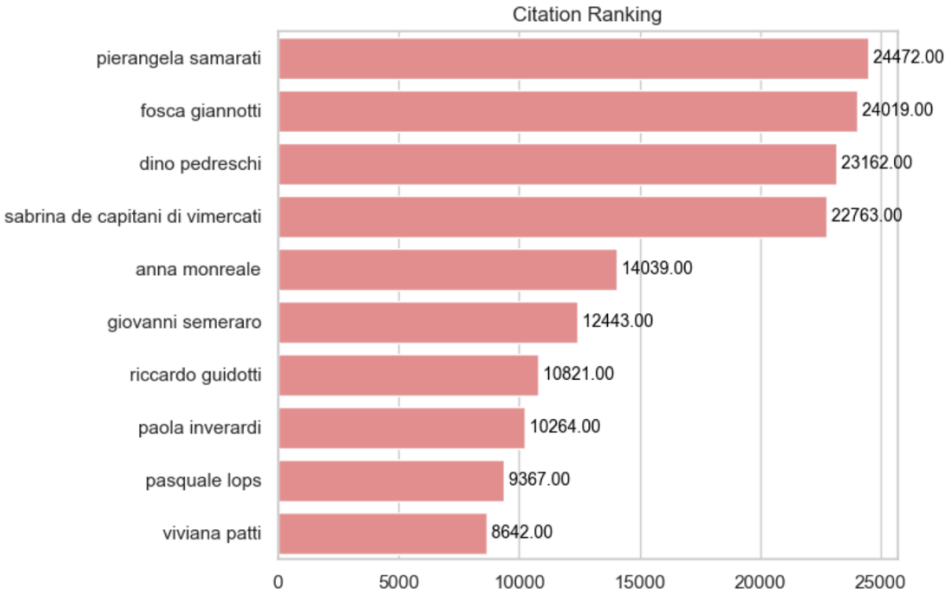

Figure 21: Citation Ranking for Computer Science

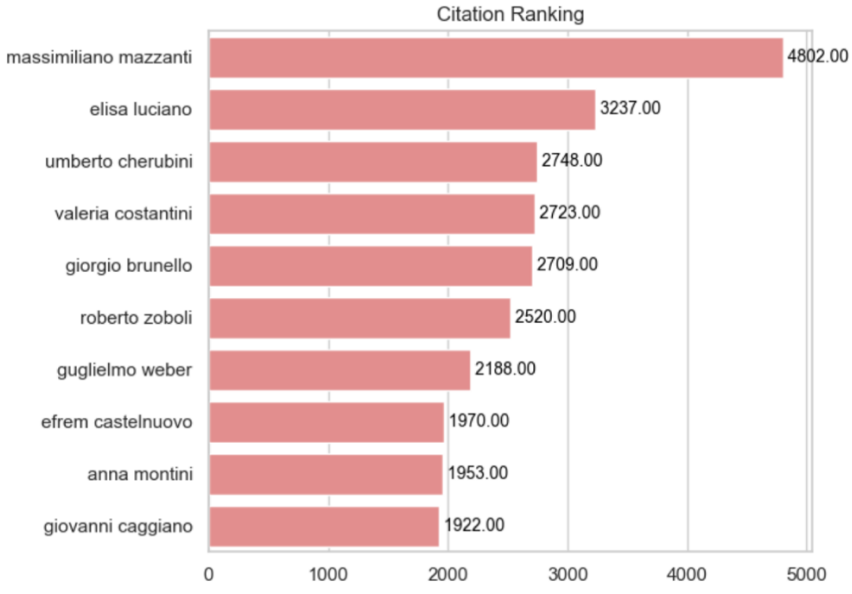

Figure 22: Citation Ranking in Economics

The h-index is a combined measure of productivity and citation impact. It indicates the number of publications (h) that have received at least h citations. In Computer Science, Pierangela Samarati has the highest h-index, indicating a strong and influential body of work. In Economics, Gianmarco Ireo Paolo Ottaviano and Massimiliano Giuseppe Marcellino share the top spot, reflecting their substantial and impactful research contributions.

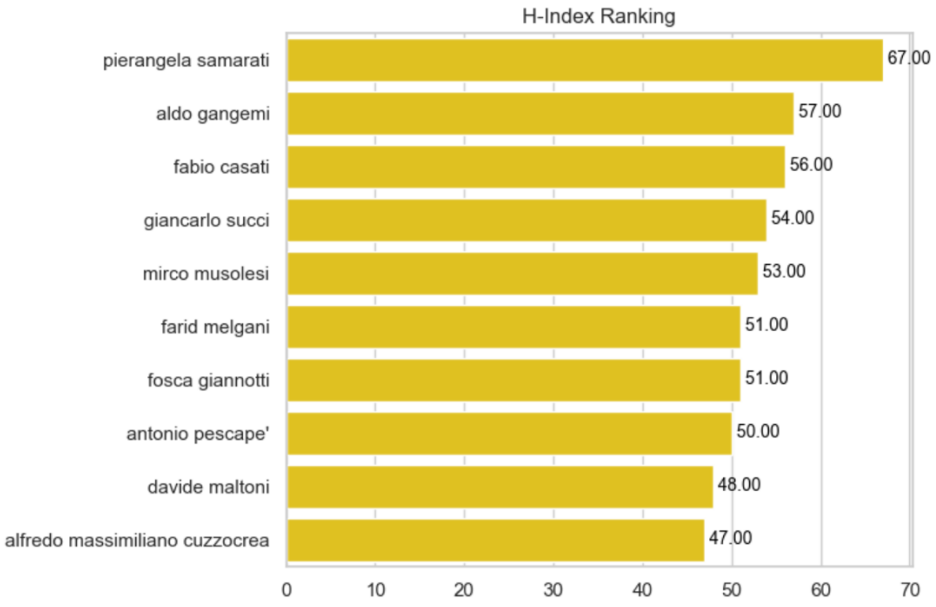

Figure 1: H-index Ranking for Computer Science

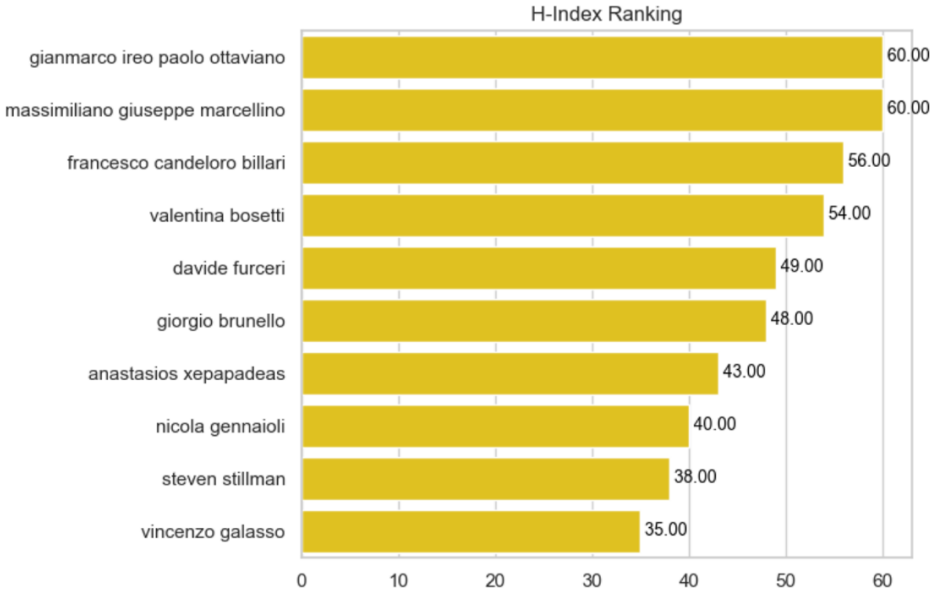

Figure 2: H-index Ranking for Economics

Combined Top Performers

In order to identify the top performers in both Computer Science and Economics, a comparative analysis was conducted across four key metrics: degree centrality, productivity, citation count, and h-index. This comprehensive approach allows for the evaluation of individual researchers who consistently perform well across multiple metrics, highlighting their overall impact and contribution to their respective fields.

The combined rankings for each author are evaluated implementing a python function "*compare_ranking*s", which splits the rankings into author names and their respective metric values. By analyzing these metrics collectively, the function identifies authors who consistently rank high across different metrics. The analysis is conducted separately for Computer Science and Economics.

Top Performers in Computer Science

The analysis reveals the top 10 performers in Computer Science who demonstrate exceptional performance across degree centrality, productivity, citation count, and h-index. These researchers are noted for their extensive collaborations, high research output, significant citation impact, and influential publications.

Top 10 Computer Science Performers:

1. Giovanni Semeraro
2. Viviana Patti
3. Alfredo Massimiliano Cuzzocrea
4. Vincenzo Loia
5. Maria De Marsico

6. Viviana Mascardi

7. Antinisca Di Marco

8. Claudio Agostino Ardagna

9. Cristina Gena

10. Aniello Castiglione

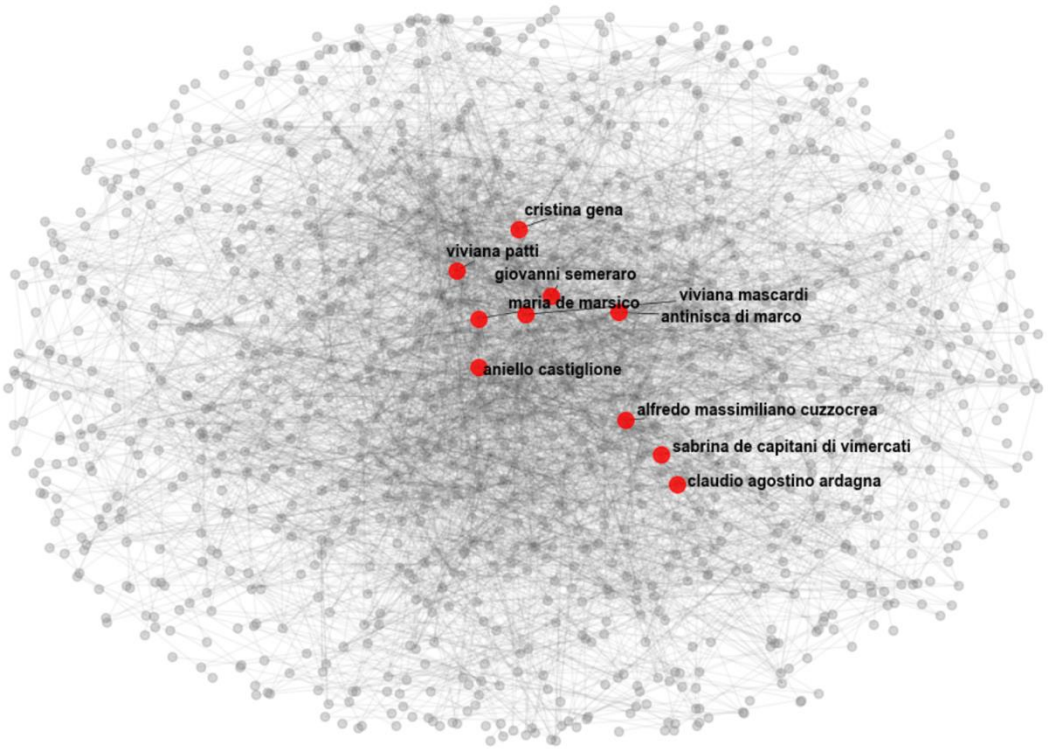

Figure 3: Top 10 Performers in Computer Science Network

Top Performers in Economics

Similarly, the top 10 performers in Economics are identified based on their consistent high rankings across the four metrics. These researchers exhibit significant collaborative networks, high productivity, substantial citation counts, and strong h-index values, underscoring their contributions to the field of Economics.

Top 10 Economics Performers:

1. Giorgio Brunello
2. Agar Brugiavini
3. Gilberto Turati
4. Guglielmo Weber
5. Paolo Buonanno
6. Luigi Grossi
7. Alessio D'Amato
8. Massimiliano Mazzanti
9. Monica Billio
10. Massimiliano Caporin

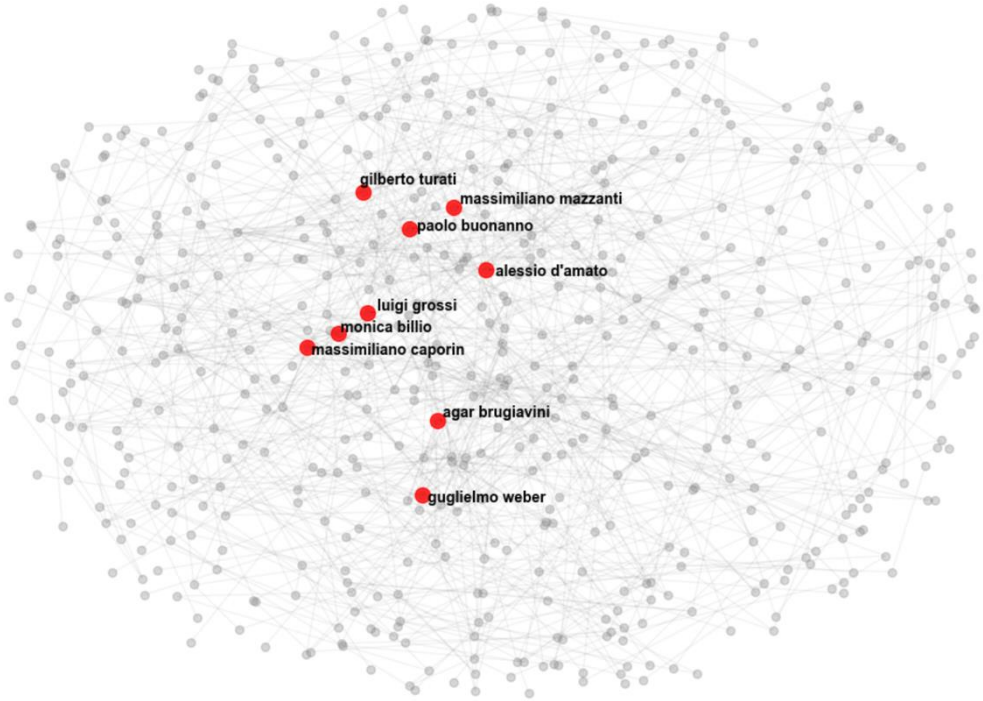

Figure 4: Top 10 Performers of Economics Network

By comparing the rankings across different metrics, we gain a holistic view of the top performers' overall impact in their respective fields. This multi-metric approach ensures that the selected top performers are not only prolific in terms of publication counts but also have a significant influence as evidenced by their citation counts and h-index values. Furthermore, the top 10 performers are also the central nodes in the graph, highlighting their key roles in the academic network through their extensive connections and collaborations.

The combined top performers in Computer Science and Economics showcase the leading researchers who excel in multiple dimensions of academic performance. These individuals are pivotal in driving the advancement of knowledge and innovation in their fields through extensive collaborations and impactful research contributions.

IV.6 – Regression Analysis

In the previous section, an extensive analysis of academic rankings was conducted based on various key metrics such as h-index, degree centrality, productivity strength, citation strength, and betweenness centrality. Top performers in each category were identified, and the correlations between these metrics were examined. Building on these findings, this section delves into the causal relationship between co-authorship and academic productivity. By leveraging advanced statistical techniques such as Propensity Score Matching (PSM) and Ordinary Least Squares (OLS) regression, this analysis aims to answer a crucial question: does co-authorship lead to higher academic productivity?

Propensity Score Matching (PSM) is employed to estimate the likelihood that a researcher has a high degree centrality based on various features (covariates). This method balances the treatment and control groups, ensuring a fair comparison between high and low centrality researchers. Propensity scores are estimated using logistic regression models, which calculate the probability of an academic having high degree centrality (a proxy for high co-authorship activity) based on the following covariates:

- Position: different academic roles, such as full professor, associate professor and researcher, which may influence productivity.

- Gender: the impact of being male (M) or female (F) on academic output.

- h-index: a metric that measures both the productivity and citation impact of publications, providing a comprehensive view of an individual's scholarly influence.

- Citation Count: the total number of citations received, indicating the recognition of one's work by peers.

- Degree Centrality: a measure of the number of direct connections an individual has in the co-authorship network, serving as a proxy for co-authorship activity.

By matching individuals on these covariates, I try to ensure that the comparison between high and low co-authorship groups is as fair and unbiased as possible.

The matching process involves pairing academics with high degree centrality (treatment group) with those who have similar propensity scores but lower centrality (control group). This method controls for confounding variables and isolates the effect of co-authorship on academic productivity.

For this process, I used the nearest neighbor matching technique, which is a commonly used technique in these analyses, where each treated individual is matched with the closest control individual in terms of propensity score.

After matching, Ordinary Least Squares (OLS) regression is conducted to examine the impact of co-authorship on academic productivity. Two dependent variables are considered in this analysis:

- Adjusted Paper Count[4] : the total number of papers published, adjusted by the number of co-authors. This metric accounts for individual contributions within collaborative works.

    - Advantages: Measures direct scholarly output and is easier to relate to the number of papers published. Useful for evaluating productivity in terms of volume of work.
    - Disadvantages: May not capture the impact or quality of the work as effectively as citation counts.

[4] By adjusting the paper and citation counts, the analysis aims to isolate the individual contribution of each researcher from the collective output of their co-authorship network. This adjustment ensures that researchers who frequently collaborate do not appear more productive solely because of the combined output of their co-authors.

- Adjusted Citation Count[3] : the total number of citations received, adjusted by the number of co-authors. This metric reflects the impact and recognition of the work within the academic community.

    - Advantages: Reflects the impact and recognition of the work within the academic community. The higher R-squared value indicates a better model fit, suggesting a more robust explanation of variance. Shows how influential the published work is, which is often a key metric in academia.
    - Disadvantages: Can be influenced by factors outside of the author's control (e.g., field trends, co-authors' networks).

Including both metrics in the thesis provides a comprehensive analysis, addressing both the quantity and the impact of academic productivity. This dual approach allows for a more nuanced understanding of how co-authorship affects different dimensions of scholarly output.

Results

Adjusted Paper Count:

The OLS regression results for adjusted paper count revealed the following:

- Intercept: Coefficient of 0.6336 ($p = 0.050$), indicating a small but significant baseline productivity.
- High Degree Centrality (Treated): Coefficient of -8.8346 ($p < 0.001$), suggesting that high co-authorship significantly decreases individual productivity by approximately 8.8 papers.
- Position (Full Professor): Coefficient of -3.4281 ($p < 0.001$), indicating that full professor positions are associated with a decrease in productivity.
- Position (Researcher): Coefficient of -0.0417 ($p = 0.856$), showing no significant effect on productivity.
- Gender (M): Coefficient of 2.0107 ($p < 0.001$), indicating that male academics have higher productivity by about 2 papers.

- H-index: Coefficient of 0.3028 ($p < 0.001$), suggesting a positive correlation between h-index and productivity.
- Citation Count: Coefficient of 0.0002 ($p < 0.001$), a positive but very small effect on productivity.

Adjusted Citation Count:

The OLS regression results for adjusted citation count revealed the following:

- Intercept: Coefficient of -175.6983 ($p < 0.001$), indicating a negative baseline level of adjusted citation count.
- High Degree Centrality (Treated): Coefficient of -239.4813 ($p < 0.001$), suggesting that high co-authorship significantly decreases citation count by approximately 239 citations.
- Position (Full Professor): Coefficient of -106.4547 ($p < 0.001$), indicating that full professor positions are associated with a significant decrease in citation count.
- Position (Researcher): Coefficient of 58.3376 ($p < 0.001$), indicating that researcher positions are associated with an increase in citation count compared to associate professors.
- Gender (M): Coefficient of 28.5823 ($p < 0.001$), indicating that male academics have higher citation counts.
- h-index: Coefficient of 13.5929 ($p < 0.001$), showing a strong positive correlation with citation count.
- Paper Count: Coefficient of 0.1894 ($p < 0.001$), indicating a positive effect on citation count.

Visual Analysis

The scatter plots of degree centrality versus adjusted productivity, both paper count and citation count, (Figure 27 and Figure 28) reveal a clear negative relationship. As degree centrality increases, indicating more extensive co-authorship, the adjusted productivity

tends to decrease. This trend is evident from the clustering of data points toward lower productivity values at higher degrees of centrality.

For adjusted paper count, the plot shows that individuals with higher degree centrality generally produce fewer adjusted papers. This suggests that extensive co-authorship may dilute individual contributions, reducing the per-author paper count. The data points are densely packed at the lower end of the adjusted paper count spectrum for higher degree centrality values, reinforcing this observation.

Similarly, for adjusted citation count, the scatter plot indicates a negative correlation. As degree centrality rises, the adjusted citation count drops significantly. This pattern implies that while these individuals may be involved in numerous collaborations, their individual impact, measured through citations, is reduced when adjusted for the number of co-authors. The clustering of data points at lower citation counts for higher degree centrality underscores this trend.

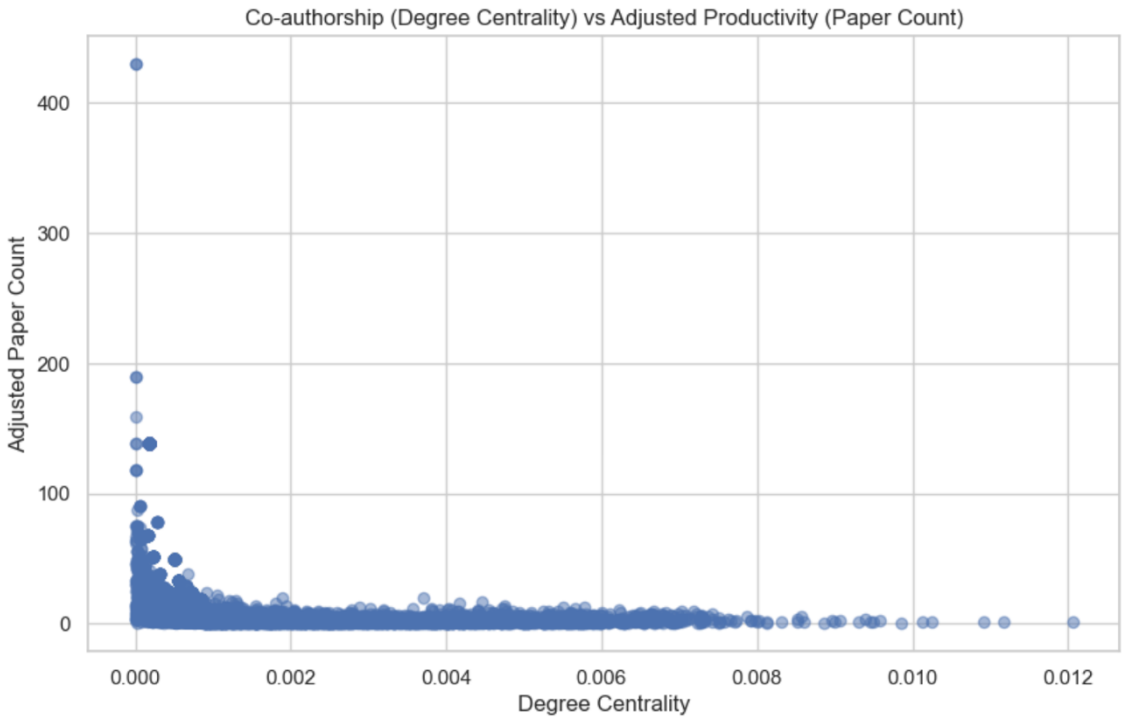

Figure 5: Scatter Plot of Degree Centrality vs Adjusted Paper Count

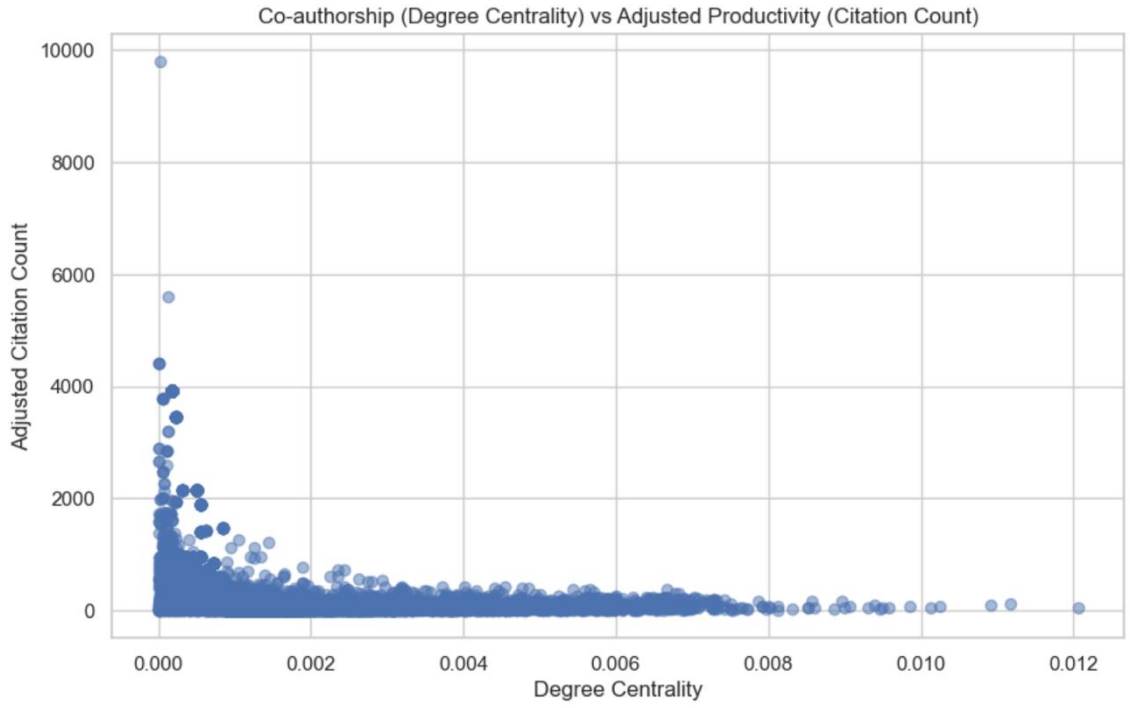

Figure 6: Scatter Plot of Degree Centrality vs Adjusted Citation Count

These visual insights corroborate the regression analysis results, highlighting the potential downsides of excessive co-authorship on individual academic productivity.

Interpretation

The analysis indicates that high co-authorship negatively impacts individual academic productivity. This suggests the presence of congestion externalities[5], where excessive collaboration may dilute individual contributions, leading to lower per-author productivity. However, the h-index and citation count positively correlate with productivity, emphasizing the importance of individual scholarly impact metrics.

[5] Congestion externalities refer to the negative effects that occur when an increase in usage of a shared resource leads to decreased efficiency or productivity for users. In the context of academia, congestion externalities may arise when high levels of co-authorship lead to diminished individual contributions. As more authors collaborate on a single paper, the individual recognition and credit for each author's work may decrease, potentially reducing their overall productivity and motivation. This phenomenon can offset the potential positive effects of collaboration, such as knowledge sharing and skill enhancement, leading to an overall negative impact on individual academic productivity.

To conclude, the findings suggest that while collaboration is essential, excessive co-authorship may reduce individual productivity. Balancing collaboration with individual research efforts could optimize academic productivity. This comprehensive analysis, using PSM and detailed regression modeling, provides robust evidence to answer the question: Does co-authorship lead to higher academic productivity? The results clearly indicate that high levels of co-authorship do not necessarily enhance productivity and may, in fact, hinder it due to congestion effect.

Chapter V – Conclusion and Recommendations

This thesis investigated the structural patterns in academic collaboration within the Italian co-authorship network, using advanced network analysis techniques to explore how researchers collaborate and the implications of these collaborations. This research examined the Italian academic landscape through various metrics such as degree distribution, clustering coefficient, and community detection, providing a detailed examination of the collaborative dynamics among Italian researchers, particularly in the fields of Computer Science and Economics.

V.1 – Key Findings

The research findings reveal that both the Computer Science and Economics co-authorship networks display small-world and scale-free characteristics. These networks are typified by a few highly connected hubs, or central nodes, that significantly enhance the dissemination of knowledge and foster substantial collaborative efforts across the network. These hubs are crucial, not just for the flow of information, but also as pivotal links that sustain the scholarly community's connectivity.

Gender and role-based disparities were prominently observed, with male researchers predominantly dominating these networks. However, female researchers, despite their fewer numbers, tend to engage more actively within their gender, suggesting a potentially supportive sub-network that could be leveraged to enhance their visibility and integration within the broader academic community. The role-based analysis underscored the importance of associate professors who facilitate substantial cross-rank collaborations, which are essential for nurturing early-career researchers and integrating them into the academic fabric.

Geographically, major cities like Rome, Milan, and Bologna emerge as central nodes. These cities not only exhibit high productivity but also maintain extensive collaborative

networks, underscoring the role of geographic proximity in shaping collaboration patterns. Proximity influences collaborative tendencies, with researchers preferring partnerships with nearby peers, which underscores the role of physical closeness in academic collaborations.

Furthermore, the study dived deep into the impact of co-authorship on academic productivity using advanced statistical techniques like Propensity Score Matching (PSM) and Ordinary Least Squares (OLS) regression. Intriguingly, the findings suggest that while co-authorship is generally seen as beneficial, there is a threshold beyond which it may lead to congestion externalities, diluting individual contributions and potentially impacting the quality and individuality of research output.

V.2 – Implications and Recommendations

These insights into the dynamics of academic collaborations carry significant implications for policy and practice within academic institutions. They necessitate a nuanced approach to fostering collaborations that are not only widespread but also equitable and efficient. Policies should particularly focus on nurturing an inclusive environment where gender disparities are addressed, and early-career researchers are supported through substantive engagement with experienced academics.

Institutions are recommended to consider the following strategies:

Balancing Collaborative and Individual Research Efforts: While collaborative research is vital, it is equally important to maintain a balance where individual efforts are not overshadowed. Policies could be crafted to encourage researchers to engage meaningfully in collaborations without overextending themselves.

Enhancing Gender Equity: The apparent gender disparities call for targeted interventions to support female researchers. This could include establishing mentorship programs, creating networks for female academics, and providing grants that specifically support women's research activities.

Supporting Early-Career Researchers: Facilitating their integration into well-established networks could significantly boost their career prospects and contribute to the overall

health of the academic ecosystem. Mentorship programs and collaborative research grants can play a critical role here.

Promoting Inter-City Collaborations: To mitigate the concentration of academic activities in major hubs, fostering collaborations that bridge the gap between major and smaller cities could encourage a more distributed and inclusive academic landscape.

Utilizing Key Influencers: Institutions should identify and leverage highly connected individuals within networks to foster broader and more integrated collaborations.

Adopting Advanced Analytical Approaches: Continued investment in developing and applying sophisticated analytical methods will be crucial for dynamically monitoring and optimizing the effectiveness of collaboration strategies.

V.3 – Concluding Thoughts

This comprehensive analysis highlights the intricacies and the critical role of collaboration in bolstering academic success. It presents a robust framework for understanding the multifaceted nature of co-authorship networks and offers practical pathways for enhancing the collaborative landscape of Italian academia. By fostering a more inclusive, balanced, and strategically interconnected research community, the potential for innovative and impactful scholarly work is significantly heightened. This study not only contributes valuable insights to the academic discourse on network analysis but also lays a foundational platform for future explorations aimed at refining and enhancing the collaborative endeavors within academia.

VI – References

1. Barabási, A. L., & Albert, R. (1999). Emergence of scaling in random networks. *Science*, 286(5439), 509-512.

2. Bianconi, G., & Barabási, A. L. (2001). Competition and multiscaling in evolving networks. *Europhysics Letters*, 54(4), 436-442.

3. Blondel, V. D., Guillaume, J. L., Lambiotte, R., & Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10), P10008.

4. Newman, M. E. J. (2001). The structure of scientific collaboration networks. *Proceedings of the National Academy of Sciences*, 98(2), 404-409.

5. Watts, D. J., & Strogatz, S. H. (1998). Collective dynamics of 'small-world' networks. *Nature*, 393(6684), 440-442.

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

7. Girvan, M., & Newman, M. E. J. (2002). Community structure in social and biological networks. *Proceedings of the National Academy of Sciences*, 99(12), 7821-7826.

8. Larson, R. C., & Wilson, R. (2011). The academic-manufacturing divide: Academic productivity in the era of technology-driven competition. *Research Policy*, 40(8), 1113-1123.

9. Moody, J. (2004). The structure of a social science collaboration network: Disciplinary cohesion from 1963 to 1999. *American Sociological Review*, 69(2), 213-238.

10. Newman, M. E. J. (2004). Co-authorship networks and patterns of scientific collaboration. *Proceedings of the National Academy of Sciences*, 101(Suppl 1), 5200-5205.

11. Wagner, C. S., & Leydesdorff, L. (2005). Network structure, self-organization, and the growth of international collaboration in science. *Research Policy*, 34(10), 1608-1618.

12. Zhang, J., & Ahn, Y. Y. (2015). Community detection in networks: Applications and algorithms. *SpringerBriefs in Complexity*, 1-85.

13. Gender disparities in academic productivity: Patterns and potential explanations. *Nature Communications*, 11, 1373.

14. King, D. A. (2004). The scientific impact of nations. *Nature*, 430(6997), 311-316.

15. Finocchi I., Martino A., Ranjbar F., Sinaimeri B. (2024). Data cleaning and enrichment through data integration: networking the Italian academia.