Study of the co-authorship network of the project Alianza EFI using graph machine learning

Carlos Garavito-Cardenas

Universidad del Rosario
Escuela de ingenieria, ciencia y tecnologia
Bogotá, Colombia
2023

Study of the co-authorship network of the project Alianza EFI using graph machine learning

Carlos Garavito-Cardenas

Thesis work presented as a partial requirement to qualify for the degree of:
Magister in applied mathematics and computer sciences

Advisor:
Ph.D. Andrés García Suaza

Research field:
Machine learning theory and applications

Universidad del Rosario
Escuela de ingenieria, ciencia y tecnologia
Bogotá, Colombia
2023

Acknowledgements

This work is dedicated to my wife, to our dreams and to our future. Thank you for your patience, love and support. Also, is dedicated to my family and friends. Thanks for your support.

Abstract

The present work demonstrates the use of graph machine learning techniques to analyze the co-authorship network among affiliated authors of the Alianza EFI Project. The document is divided into three chapters: the first one provides a comprehensive overview of the global and local context of Artificial Intelligence (AI) in a way that justifies the significance of working with AI topics in today's world. The second chapter is dedicated to constructing the theoretical framework for working with graphs and machine learning. The final chapter showcases the results of implementing graph machine learning for predictive tasks at the node, link, and community levels. Specifically, this chapter reveals that the Alianza EFI project involves contributions from 390 unique authors, associated with 112 distinct institutions, resulting in 274 unique products. It also demonstrates that the Universidad del Rosario plays a central role in institutional collaborations, in contrast to the other institutions within the alliance. Finally, after applying graph machine learning techniques, it was observed that these strategies enable the alliance to identify new research topics for authors, establish new connections among isolated authors, and discover new communities of research interests.

Key words
Machine learning – Deep Learning – Graphs – Graph Machine Learning

Problem Statement and Justification

In the Colombian context, the generation, promotion, and maintenance of scientific research networks have a significant impact on the country's science and technology development. These networks foster cooperation among educational institutions, researchers, and industrial sectors, leading to the creation of shared knowledge that addresses local challenges. One effective way to comprehend the dynamics of these research networks is through the use of bilbiometric analysis. Essentially, a citation network graphically represents links between various academic documents such as papers or books. It enables the analysis of interactions between these documents and their authors, as well as the mutual impact they have on each other through citations.

Acording to [18], a citation network comprises nodes representing documents and links representing citations between them. Citation networks are potent tools for bibliometric analysis, allowing the exploration of influence patterns between authors and topic connections. They also facilitate knowledge dissemination within scientific communities. These networks prove valuable for identifying exceptional papers, influential authors, emerging research trends, and novel research topics.

Citation networks hold interdisciplinary significance, attracting interest from fields such as linguistics, computer science, and mathematics. Publications related to citation networks date back to 1972, with [11], where the concept of using citations as a measure of academic influence and importance was introduced. In recent times, with the surge of machine learning, novel computational techniques have been employed to advance citation network analysis. For instance, [23] provides a comprehensive overview of applying deep learning methods to solve machine learning tasks on graph data. It shows how Graph Neural Networks (GNNs) have emerged as a powerful tool for analyzing and learning from graph-structured data, including social networks, molecular structures, and citation networks. The authors categorize GNNs into different classes, including graph convolutional networks, graph attention networks, and graph autoencoders and discusses how GNNs have been applied to address complex tasks like node classification, link prediction, and graph classification.

As previously established, a citation network serves as a powerful tool for understanding interactions and influences among academic documents, authors, and institutions. In the context of the Alianza EFI, where collaboration spans across numerous institutions and experts, the intricate web of connections and knowledge exchange is intricate.

The Alianza Economía Formal e Inclusiva (EFI) is an alliance established and chosen through the Colombia Científica initiative. It proposed a scientific program titled "Inclusión productiva y social: programas y políticas para la promoción de una economía formal" (Productive and Social Inclusion: Programs and Policies for Promoting a Formal Economy), which secured resources of 18 thousand million Colombian pesos for implementation over the next

four years. The initiative is administered by the Universidad del Rosario, serving as the anchor institution of the alliance.

The primary objective of the EFI project is to diagnose, examine, and address factors and barriers that impact the social and productive inclusion of economic agents across diverse contexts from a systemic perspective. This is achieved through qualitative and quantitative studies conducted within a robust ecosystem involving various actors and economic and social sectors. These studies encompass diagnostics, intervention designs, and evaluations aimed at promoting the social and productive inclusion of economic agents. The focus is on populations and geographic areas traditionally excluded from the formal economic system.

Colombia Científica is a government program led by the Ministry of Education, Ministry of Commerce, Industry, and Tourism, Colciencias, and Icetex, with support from the World Bank. The program's goal is to encourage scientific projects in Colombia and improve the quality of higher education institutions. It fosters research and innovation projects that contribute to regional development and address the needs of the productive sector.

To achieve its objectives, the EFI alliance comprises collaborations between accredited and non-accredited higher education institutions, entities involved in productive sectors, businesses, and international partners. These collaborations are led by an accredited "anchor" university institution.

The EFI alliance carries out its publications under the framework of seven projects: Entrepreneurship, Understanding Inclusive Labor Markets, Rural Economics, Cities as Scenarios for Social Inclusion, Macro Institutional Aspects, Understanding the Informal Economic Agent, and a Social Laboratory.

Hence, the benefits of study a citation network in the context of a project that involves multiple institutions and authors like the Alianza EFI, can have significant benefits:

Identify influences and connections Analyzing a citation network allows for the identification of influences and connections between documents, authors, and institutions. This can encourage more frequent collaborations and reveal new topics of interest, thereby facilitating a deeper understanding of the research dynamics within the project.

Impact evaluation The citation network can be used to evaluate the impact of documents, authors, and institutions. It allows the identification of authors who have a greater influence on the project.

Detection of topics communities The analysis of communities within the citation network can assist in grouping authors according to topics. This could prove valuable for interdisciplinary projects, as it allows for the identification of how different disciplines contribute to the project and how they interweave among one another.

Collaborations visualization By utilizing the analysis of the citation network, it becomes possible to visualize collaborations among various authors and institutions. This capability aids in identifying patterns of successful collaboration, thereby facilitating the discovery of new opportunities for associations.

Identification of research gaps By identifying research topics or projects that have received less attention in terms of production, author contributions, or institutional contributions, it is possible to discover new opportunities to address research gaps within the project. This approach can lead to the strengthening of these gaps, ultimately generating new knowledge that contributes to the project's objectives.

Support for decision making The analysis of the citation network could provide valuable information for decision-making within the project. It will help identify key results, key authors, or key institutions that should be promoted or prioritized for future research papers.

In essence, the citation network analysis applied to the Alianza EFI project could bring a more complete vision about the collaborations, authors and institutions influences and key results of the project, which could enrich the decision making to enhance the impact of the project on the academic community.

Co-authorship analysis holds the potential to continue playing a pivotal role in understanding and supporting the Alianza (EFI) project, even beyond its official funding period. While the project's official financial support may come to an end, the network of collaboration and knowledge exchange that has been cultivated within the alliance remains a valuable asset. By leveraging co-authorship analysis techniques, the alliance can continue to gain insights into the evolving dynamics of its network, identify potential areas for collaboration, and uncover emerging research trends. This ongoing analysis can serve as a guiding compass for maintaining the network's vibrancy and relevance in future scenarios. Understanding the patterns of co-authorship and collaboration that have emerged throughout the project's duration will enable the alliance to foster connections, encourage interdisciplinary interactions, and facilitate knowledge dissemination among its members. Therefore, co-authorship analysis offers a proactive approach to ensuring the sustainability and growth of the network, even in the absence of official funding, by empowering the alliance to make informed decisions and cultivate a thriving ecosystem of research collaboration and knowledge exchange.

One of the significant outcomes of the Alianza EFI project is the comprehension of establishing an ecosystem and a scientific agenda focusing on informality, which involves numerous institutions and experts. Hence, the problem statement for this project is as follows:

> How to gain a comprehensive understanding of the construction of a robust scientific agenda concerning a topic that encompasses multiple authors and institutions?

Work Objectives

General objective

Employ graph learning algorithms on the co-authorship network of the Alianza EFI alliance to comprehend the creation and evolution of an extensive scientific agenda revolving around specific topics. These topics involve numerous authors and institutions. This approach aims to facilitate the understanding and prediction of the impact of individual authors, their collaborations, and the communities centered around these topics.

Specific objectives

- Conduct a comprehensive descriptive analysis of academic papers, including examination of authors and affiliated institutions.
- Employ graph learning algorithms to attain accurate node classification within the network.
- Utilise graph learning algorithms to predict links effectively in the network.
- Apply graph learning algorithms to successfully identify communities within the network structure.

Content

1 ARTIFICIAL INTELLIGENCE CONTEXT

Artificial intelligence (AI) is no longer a topic confined to science fiction; rather, it has become a practical technological tool with the potential to revolutionize numerous aspects of society. In this regard, the year 2022 witnessed the deployment of a vast array of AI tools, including noteworthy mentions like ChatGPT and DALL-E. To establish a contextual framework and avoid delving into philosophical questions, within the scope of this work, AI refers to the capacity of machines to execute tasks without human supervision, leveraging computational power [20]. As mentioned earlier, there exists a dynamic evolution in the field of AI, making it essential to comprehend its impacts on various societal dimensions. To achieve this understanding, numerous researchers and reports attempt to gauge the progress of this field; among them, one succinctly captures the state of the art. This report is the AI Index Report, led by Stanford University and published annually. It offers an overview of the current research, development, and adoption of AI.

Despite the fact that the development of AI has seen strong contributions from developed regions like the United States, China, and Europe, Colombia cannot be isolated from this discussion, and its unique context must also be analyzed. Therefore, the Colombian government is conscious of the impact of this new technology and how it can affect the local context. It has consequently developed AI legislation to adopt AI for the benefit of society. In particular, to assess the state of the art of AI development in Colombia, a group of experts known as the AI Experts Mission was established. This group aims to provide recommendations on how Colombia should adopt AI. Consequently, the purpose of this chapter is to summarize key findings from the AI Index Report and compare them with the AI Experts Mission's conclusions related to Colombian society.

1.1 An overview of the global context of AI

The AI Index report covers an extensive array of topics associated with artificial intelligence, encompassing research and development, technical performance, AI ethics, the economy, education, policy and governance, diversity, and public opinion. Although all these facets hold significance, this section will concentrate on specific findings concerning global AI in research and development, the economy, education, and policy and governance.

Research and development Observing the quantity of AI publications, it becomes evident that there is a nonlinear surge in the volume of published works. Figure **1-1** illustrates the total global publications spanning from 2010 to 2021. It is noticeable that the number of publications in 2021 was more than twice the count compared to a decade prior.

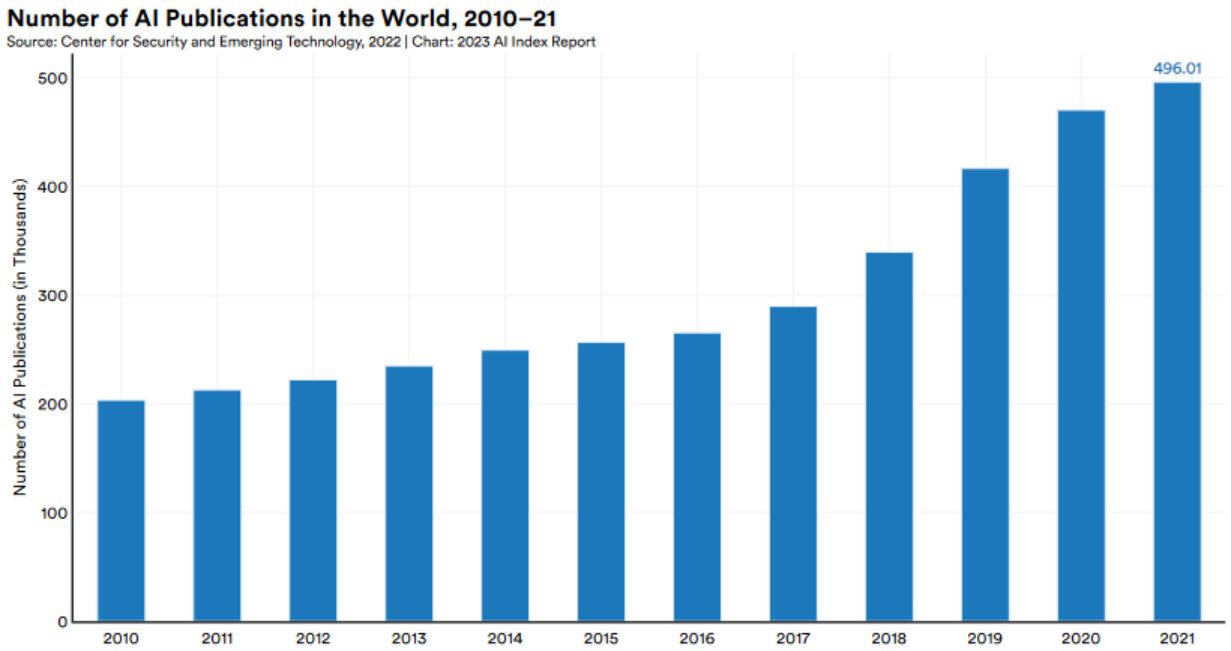


Figure 1-1: Number of AI publications in the world, from 2010 to 2021. Taken from [17]

Upon closer examination of the fields of study, Figure **1-2** illustrates the progression in the number of publications across various disciplines. It can be observed that pattern recognition garners the primary focus of research, comprising 26% of the total publications. It is closely followed by machine learning at 20% and computer vision at 13% of the overall contributions. Notably, pattern recognition has experienced a nearly threefold increase in publications between 2011 and 2021. Moreover, it is evident that machine learning has exhibited significant growth, particularly from 2017 onward, making it the fastest-growing field compared to others. Lastly, it is worth mentioning that natural language processing contributes 7%, which is relatively low compared to other topics, yet it remains one of the most accessible applications for non-technical audiences, as exemplified by ChatGPT.

In terms of cross-country collaboration, the main contribution is made by the work of the United States and China, where, despite the political tensions related to espionage allegations, for the 2021 year reaches 10.470 publications [17]. At the same time, China also reaches an important amount of publications in collaboration with the United Kingdom, reaching to 4.130 publications. In the third place of collaborations, it is possible to find United States and Germany with 3.420 publications.

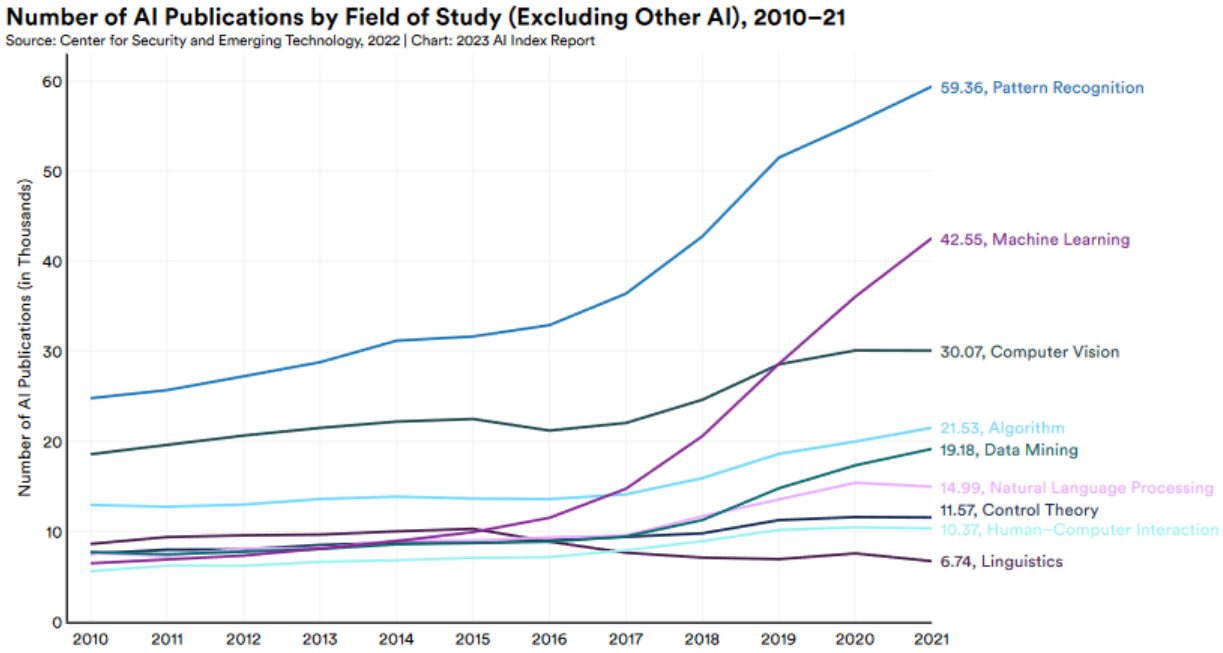


Figure 1-2: Number of AI publications by field of study, from 2010 to 2021. Taken from [17]

To comprehend notable trends in machine learning systems, the AI Index report analyzes information compiled by Epoch AI, a consortium of researchers who curate a database containing details about significant AI and machine learning models. The criteria for model inclusion encompass contributions to the cutting edge, historical significance, and a substantial citation count [17].

For discerning contributions by country, the Epoch AI team tallies the number of significant machine learning systems produced by each country since 2002. As depicted in Figure **1-3**, the preeminent contribution hails from the United States, trailed by Canada and subsequently China. Notably, Latin America and the Caribbean region see contributions from Mexico and Argentina, yet in each case, their contributions do not surpass 10 significant systems.

Economy Starting in the year 2014, the United States has exhibited a consistent growth pattern in terms of its job postings. In fact, since 2016, it has held the position of the country with the highest number of job offerings. As of 2022, the United States remains at the forefront with the highest percentage of job postings, standing at 2.05%. Following closely are Canada, contributing 1.45%, and Spain with 1.33%.

In a similar vein, when examining the sought-after skills in AI job postings within the United States, Figure **1-5** illustrates that the top three most highly demanded skills in 2022 are Python, accounting for 37.13%, followed by computer science at 32.58%, and SQL with

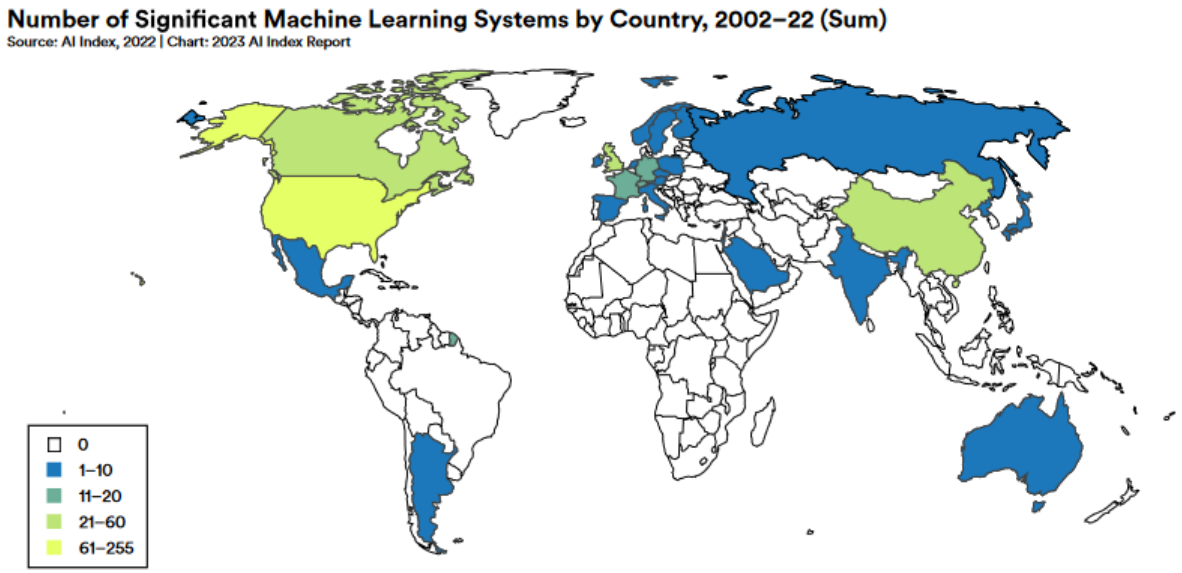


Figure 1-3: Number of significant machine learning systems by country, from 2002 to 2022. Taken from [17]

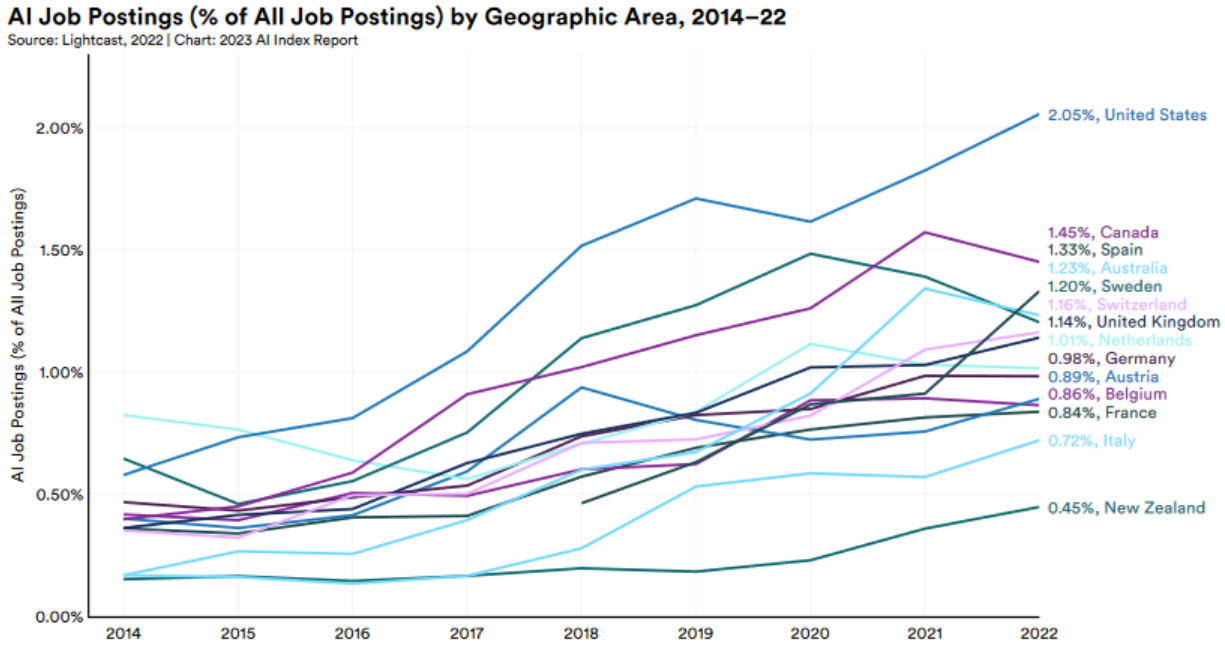


Figure 1-4: AI jobs postings by country, from 2010 to 2022. Taken from [17]

23.25%. It is noteworthy that in comparison to the past decade, the demand for Python has surged by 592%, while computer science has experienced a 63% rise, and SQL has seen a 153% increase.

Additionally, it is worth observing that the second skill with the most substantial demand

growth is Agile methodology, showing a remarkable 509% increase. This indicates that for companies in the United States, the significance of skills extends beyond the technical realm, encompassing the methodologies employed for project development.

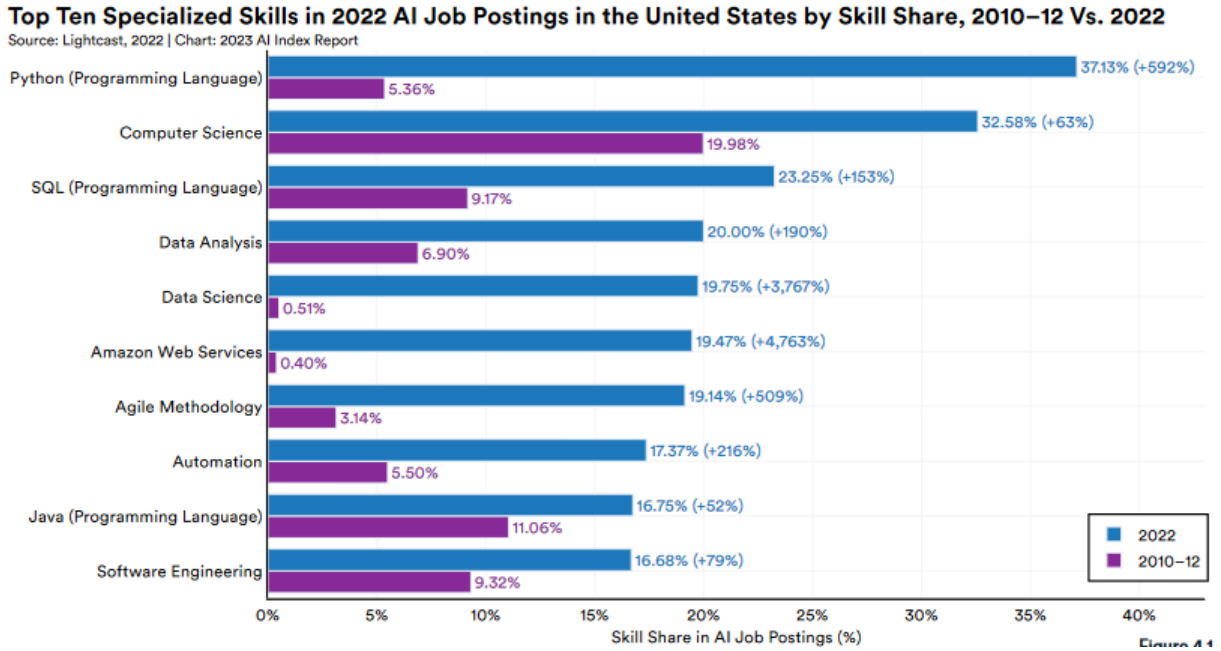


Figure 1-5: Publications

To observe the temporal progression of private investment in AI by region, refer to Figure **1-6**, which indicates that the United States has maintained the highest level of investment since 2013. Additionally, the rate of investment growth has accelerated after 2016. Notably, the pinnacle of investment was reached in the year 2021, with an impressive sum approaching $70 billion USD. It's important to highlight that even though the United States experienced a 41% decrease in investment in 2022 compared to 2021, it still remains the leading investing region by a significant margin, surpassing both China and the European Union..

Policy and governance According to [17], between 2016 and 2022, the United States has enacted a total of 22 AI-related legislations, securing its position at the forefront. Portugal and Spain occupy the second and third ranks with 13 and 10 legislations respectively. Notably, in the case of the United States, the trajectory of AI-related legislative actions began to surge in 2016. In that year, there were no laws specifically addressing AI-related concerns. However, by 2021, an impressive 134 bills related to AI had been introduced. Up to this point, 2021 remains the year with the highest number of proposed AI bills. In 2022, 88 new bills were introduced, of which only 9 progressed into law [17]. This shift underscores the growing interest within American society to establish legal frameworks that accommodate emerging AI technologies.

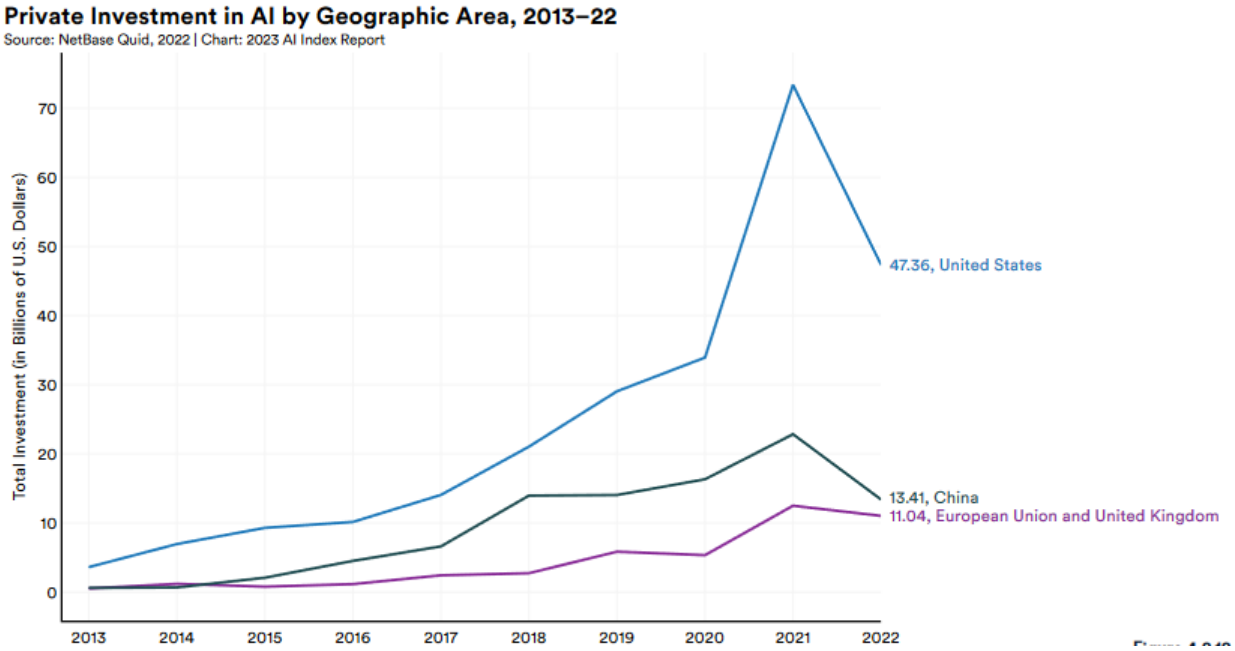


Figure 1-6: Private investment in AI by geographic area, from 2013 to 2022. Taken from [17]

Conversely, it's important to acknowledge that within the Latin America and Caribbean region, AI-related legislations are predominantly confined to Brazil and Argentina.

Conversely, in contrast to legislation, numerous countries have formulated national strategies for the integration of AI. As outlined in [17], in this regard, the Latin America and Caribbean region demonstrates a significant involvement. It is noteworthy that among the countries in this region, only Venezuela, Ecuador, Guyana, Suriname, Bolivia, and Paraguay have yet to publish any AI-related strategies.

In the specific case of Colombia, the roadmap for AI adoption is outlined in CONPES documents, as well as through insights and recommendations provided by the AI Expert Mission. The upcoming subsection will delve into this aspect in detail.

1.2 An overview of the local context of AI

Colombia possesses a significant opportunity to devise strategies aimed at adopting and implementing AI, thereby ensuring preparedness to address both societal and market demands, and fostering competitiveness within regional and global markets. The nation boasts a pool of talented individuals, yet it is imperative to establish policies that facilitate widespread access to learning, training, and employment opportunities in the field of AI. The ensuing section delves into crucial facets associated with Policy and Governance, Economy, as well as Research and Development within the realm of AI in the country.

Policy and governance In 2019, the Colombian government formulated its policy for digital transformation and the adoption of artificial intelligence through the document known as CONPES 3975. This policy aims to bridge the information gaps concerning AI technologies between the public and private sectors, while ensuring the academic sector's access to cutting-edge international tools and knowledge related to AI. According to the Global AI Index compiled by Tortoise, Colombia ranks 48th globally among countries involved in AI development and adoption through investment, innovation, and implementation. One significant finding of this report indicates that Colombia faces a shortage of trained personnel in specialized domains like data science, artificial intelligence engineering, and data engineering [22]. Recognizing the imperative to bolster Colombia's capabilities in AI technologies, the government has taken steps to establish a team of experts tasked with formulating a medium and long-term strategy to embrace AI as a developmental tool for Colombian society.

The current status of Colombian public policies pertaining to AI matters is encapsulated in the following documents: CONPES 3920, outlining policies for data utilization and exploration employing big data technologies; CONPES 3975, establishing 14 guiding principles for AI development in Colombia with the goal of fostering societal and economic advantages across both public and private sectors; and CONPES 4023, underscoring the necessity to fortify digital talent capabilities as an economic catalyst for post-COVID crisis recovery [22].

Economy In terms of employment and the requisite expertise necessary for effective AI implementation, [22] highlights that Colombia currently lacks the skilled workforce essential for utilizing and integrating AI technologies. Moreover, it underscores that due to Colombia's significant inequality, the population most vulnerable to the impact of AI consists of those with limited resources. Their limited access to and training in these technologies exacerbate this disparity. This assertion finds support in Colombia's performance on several international AI assessments. Notably, the 'AI Readiness and Inclusion Index' by Microsoft predicts that by 2030, Colombia will need to elevate its highly qualified talent pool from 17% to 50% in relation to the total workforce to meet market demands. In the 'Government AI Readiness Index' by Oxford Insights, which evaluates metrics such as the number of STEM graduates, the quality of graduate education, the adoption of digital skills, and the availability of 'deep-knowledge' jobs, Colombia scored 47.4 out of 100. Another assessment, the 'Global Talent Competitiveness Index 2020' by INSEAD, which assesses the capacity to cultivate, attract, and retain international talent while bridging knowledge gaps, positions Colombia at 74th place out of 132 countries. Lastly, the World Bank's report 'Going Viral: COVID-19 and the Accelerated Transformation of Jobs in Latin America and the Caribbean' highlights that 48% of Colombia's current jobs are at risk due to automation.

In the realm of investment, Figure **1-7** portrays a remarkably intriguing pattern. It is evident that there has been a substantial upsurge in investment, exemplified by the notable difference between 2022 and 2016, where the investment has grown by over 10 times.

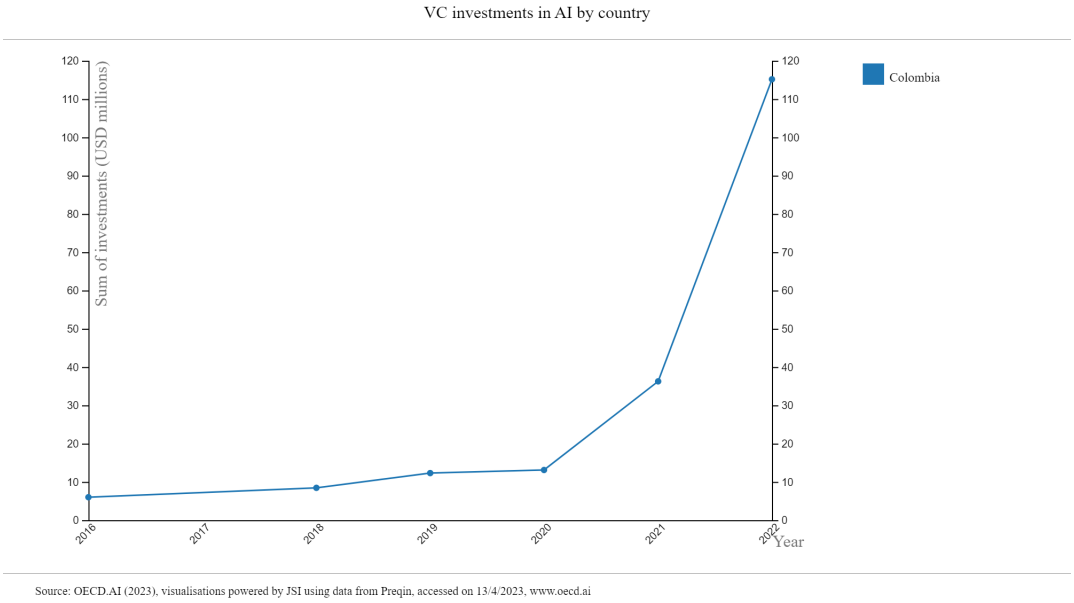


Figure 1-7: Venture capital investment in AI in Colombia. Taken from [1]

Research and development Regarding AI research conducted by educational institutions, Figure **1-8** illustrates a significant trend. The institution that produces the highest number of scientific papers is the National University of Colombia, exhibiting a substantial lead from 2000 to 2022 compared to other institutions. In the year 2022, the three most prolific institutions are the National University with approximately 185 publications, followed by the University of Antioquia with 100 publications, and subsequently the Pontificia Universidad Javeriana with 80 papers.

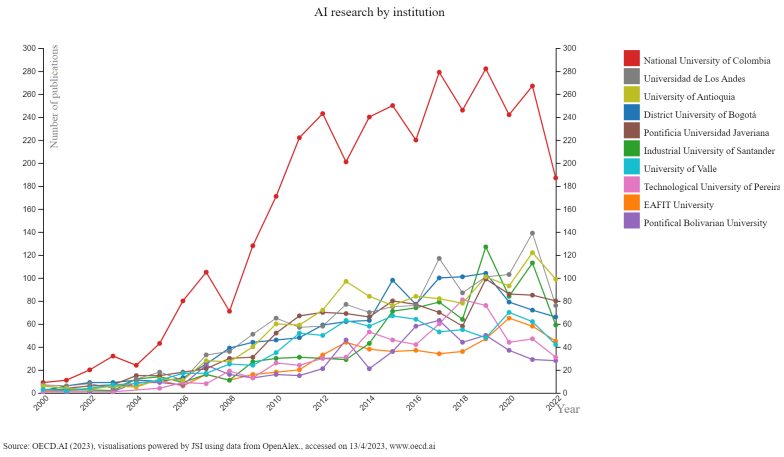


Figure 1-8: AI research by institution. Taken from [1]

Therefore, with these results, the expert mission on AI recommends the following strategies to create a sustainable framework for the adoption and implementation of AI in Colombia

[22]:

1. Count with a focus on inclusion and empowerment,
2. Develop AI talent programs, recognising the skills diversity of the country,
3. An active observation to the generation of initiatives in talent and future of the work,
4. Promote a AI future vision and enhance the knowledge generation,
5. Promote the use of technology with a non oblivious vision of society and the proper context of the country,
6. Identify specific applications of AI in Colombia to promote the development of capabilities,
7. Prioritise the analysis and implementation of the recommendations of environment sustainability and AI,
8. Contribute to the implementation of the ethical frame of AI in Colombia,
9. Integrate Colombia in an sustainable way in the global knowledge flows. Guarantee the permanent access to the population and to the productive apparatus to new technologies.

Conclusions

Artificial intelligence is a topic of burgeoning interest worldwide, captivating both academia and industry. Predominantly, key publications thrive within scientific journals, with the foremost research topics encompassing pattern recognition, machine learning, and computer vision. Remarkably, China emerges as the pacesetter in terms of publication output, and nine out of the top ten global research institutions hail from this nation. In terms of the economy, the United States stands as the dominant contributor to the total job postings, accounting for 2%, with Python, computer science, and SQL ranking as the most sought-after skills.

On a local scale, the expert mission unveiled Colombia's substantial journey ahead in terms of policies and legislation essential for adopting AI as a tool for societal transformation. Within the realm of publications, the National University of Colombia claims the lead, boasting the highest number of publications from 2000 to 2022. Turning to investment, Colombia's resolute commitment is evident through its progressive investment endeavors, with investments increasing over tenfold compared to the previous decade. Aligned with an acute awareness of both the prospects and challenges, the expert mission proposes a comprehensive nine-point framework designed to utilize AI as a driving force for societal metamorphosis.

2 GRAPH MACHINE LEARNING

Artificial Intelligence (AI) and Machine Learning (ML) constitute disruptive fields that are ushering transformative changes into our society. Despite often using AI and ML interchangeably, a conceptual distinction exists between them. AI can be comprehended as a suite of technologies geared towards resolving tasks without human supervision, whereas ML is merely one technique employed to achieve this objective. ML, an expansive domain, can be categorized into two primary types: classical machine learning and deep learning. Within classical ML, common algorithms for classification and regression include logistic regression and decision trees, among others. Deep Learning (DL) encompasses a vast array of applications utilizing neural networks, including Convolutional Neural Networks (CNNs) and Generative Adversarial Networks (GANs) [5], among others. Notably, consideration of data nature, whether structured or unstructured, remains pivotal.

Structured data, often presented in a row-and-column format, lends itself to direct manipulation through classical ML algorithms. Conversely, unstructured data like text, images, audio, or video can be transformed into structured data using mathematical constructs to facilitate the application of DL algorithms. Nonetheless, complex data may resist structuring yet still adopt a graphical or network representation. This applies to scenarios such as social network interactions, network communications, banking transactions, or molecular structures. Consequently, ML challenges involving complex data structures necessitate potent techniques to infer graph properties, enabling them to be harnessed by ML algorithms. The discipline focused on handling input data in the form of graphs is aptly named graph learning [24]. Therefore, this chapter expounds upon the nature of graph learning, the range of problems amenable to its application, and presents a mathematical foundation for graph manipulation. It further outlines the primary strategies for executing node embedding, together with a survey of the most prevalent algorithms utilized for ML at the levels of nodes, links, and entire graphs.

2.1 Introduction to machine learning

AI and ML are terms frequently conflated, yet they differ significantly from a theoretical standpoint. As per [20], AI pertains to constructing machines capable of proficient and secure performance across a broad spectrum of scenarios. While multiple interpretations of AI exist, this work subscribes to *The Rational Agent Approach* expounded in [20]. For

instance, envision a self-driving car: here, the car functions as an intelligent agent, and the road represents the environment with which this agent engages. Depicting this notion, Figure **2-1** delineates the agent's interaction with its surroundings via sensors and actuators. Notably, a question mark within the agent symbolizes the process of the agent scrutinizing sensory data to formulate responses executed by its actuators; this constitutes the algorithmic core. Consequently, AI embodies the symbiotic fusion of electronic, electrical, or mechanical systems and computational processes.

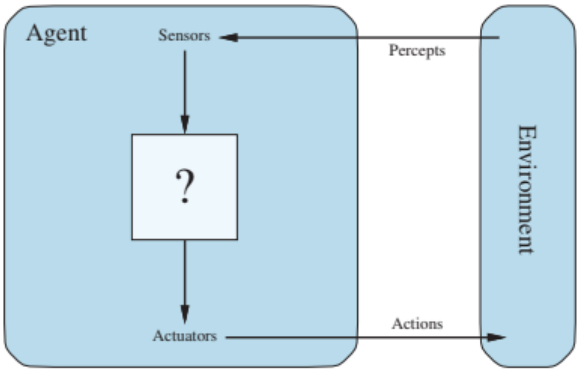


Figure 2-1: Scheme of an intelligent agent that interacts with its environment through sensors and actuators. Taken from [20]

Keeping this perspective in mind, Machine Learning (ML) can be regarded as a subset of AI. Specifically, ML embodies the facet of AI that facilitates decision-making and enhances performance for agents. Consequently, ML occupies a multidisciplinary realm and finds application in diverse AI tasks such as computer vision, natural language processing (NLP), and robotics.

Moreover, irrespective of the task at hand, ML can be classified into two categories: classical machine learning and deep learning (DL). A principal distinction between classical machine learning and DL methods lies in their performance characteristics. Classical methods exhibit a performance plateau as the dataset size increases, displaying minimal improvement. In contrast, DL methods exhibit an inherent capacity to capitalize on larger datasets, continuously enhancing their performance alongside data augmentation. This phenomenon is illustrated in Figure **2-2**."

2.2 Introduction to machine learning with graphs

The roots of graph theory trace back to Euler in the year 1735 when he delved into the study of the seven bridges of Königsberg. This particular problem stands out as Euler adeptly resolved it using graph representations, assigning nodes to land masses and edges to bridges. Such challenges belong to the realm of complex systems. The remarkable aspect of working

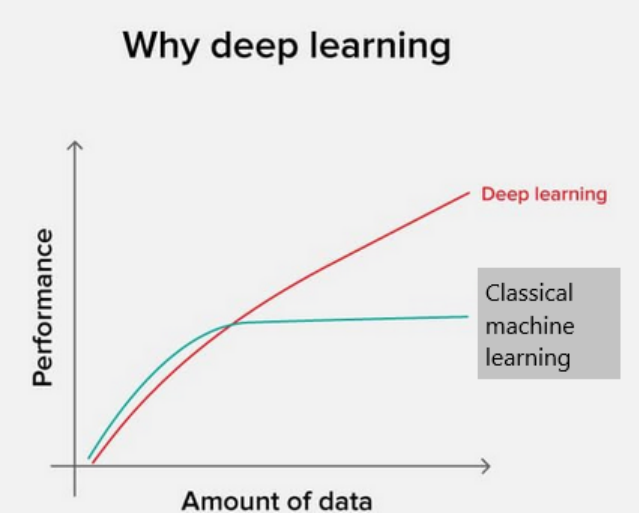


Figure 2-2: Comparison of the performance of deep learning and classical machine learning methods in function of the amount of data. Modified from [19]

with graphs, or networks, lies in their versatility. The same graph representation can be employed across multiple problems, whereby a solution for one problem can be extrapolated to others as long as the underlying graph structure remains consistent. This concept finds clarity in Figure **2-3**, where subfigure *a* depicts a computer network, *b* illustrates character relationships in the movie 'The Godfather,' *c* outlines protein interactions, and *d* presents the overarching graph representation encompassing all cases [8].

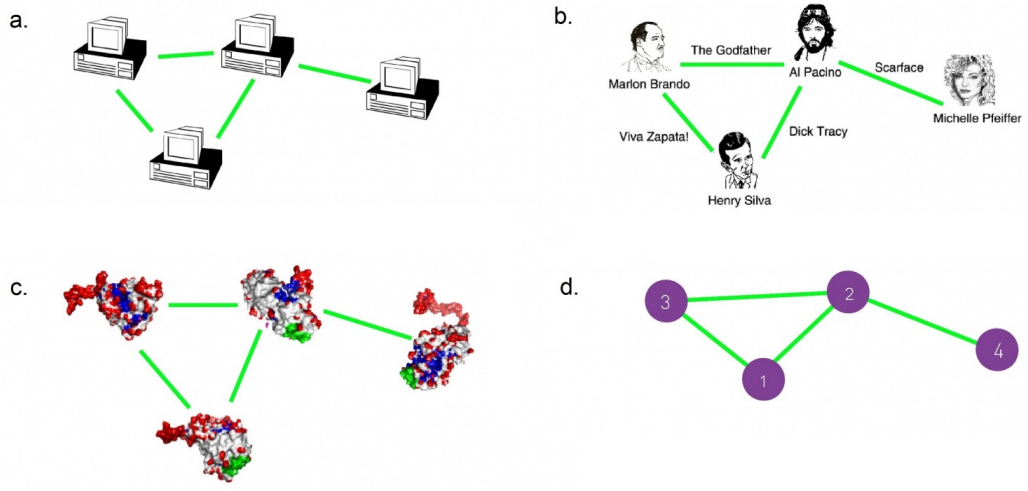


Figure 2-3: Graph representation of different kind of problems but with the same graph structure [8].

The distinction between structured and unstructured data is well-established. In the former,

data assumes a tabular form characterized by rows and columns akin to tables. Conversely, unstructured data, including audio, images, and text, defies such tabular representation. Among unstructured data, graphs stand out as a distinct type, often utilized to depict interconnected information. Renowned applications of graph data structures span a spectrum encompassing social network analysis, financial transaction systems, biological networks, transportation systems, and telecommunication networks, among others [10]. Thus, comprehending methodologies capable of discerning patterns inherent to graph structures and extrapolating the encoded information becomes pivotal.

Particularly noteworthy are deep learning techniques that effectively process graph data. Within this purview, artificial neural networks dedicated to this task are dubbed graph neural networks (GNNs) [13]. Broadly, the domain of machine learning dealing with graph data assumes the nomenclature of graph machine learning [24].

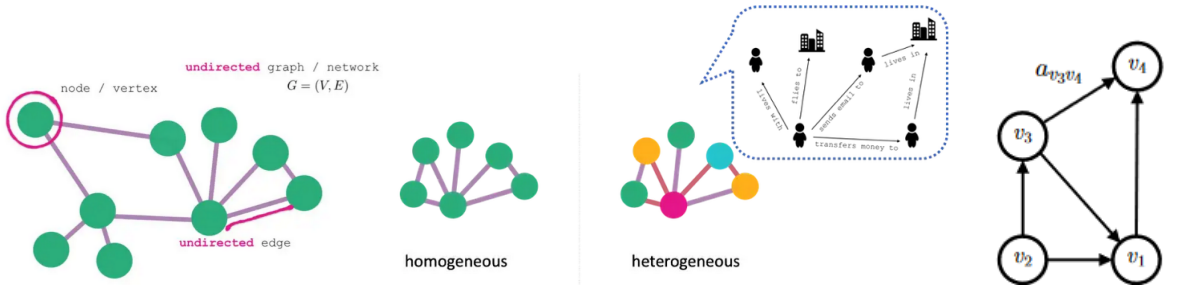
In the ensuing sections, an in-depth exploration is undertaken, delineating the essence of graphs, elucidating the gamut of problems amenable to graph machine learning, and culminating in an exposition of its application to the unique domain of citation networks.

What is a graph?

As previously highlighted, graphs serve as unstructured data representations utilized for explicating complex systems [12]. Graphs are also denoted as networks, with the distinction between the two residing solely in their nomenclature for components. A graph essentially comprises two fundamental constituents: nodes and edges. Nodes correspond to entities, while edges signify the relationships linking these entities. In network context, nodes are termed vertices and links correspond to edges. Throughout this document, graph terminology will be employed. A rudimentary depiction of a graph is featured in Figure **2-4**a, where green circles depict nodes and purple lines delineate edges. Figure **2-4**b underscores the existence of two graph types: homogeneous graphs and heterogeneous graphs. Homogeneous graphs entail nodes and links endowed with singular attributes, while heterogeneous graphs accommodate diverse attributes for nodes and links. For instance, in the heterogeneous graph illustration, node color could signify a person's role in a company, and link colors could denote various interactions such as email communication, monetary transfers, collaborations, and more. Lastly, Figure **2-4**c illustrates the concept of directed edges, wherein a specific direction governs the connection between edges.

Formally, a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ is defined by a set of nodes $\mathcal{V}$ and a set of edges $\mathcal{E}$ connecting the nodes. An edge going from node $u \in \mathcal{V}$ to node $v \in \mathcal{V}$ is denoted $(u, v) \in \mathcal{E}$. Edges are undirected if are equal in both directions, this is, if $(u, v) \in \mathcal{E} = (v, u) \in \mathcal{E}$ [12]. In addition, a heterogeneous graph is denoted as $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathcal{X}, \mathcal{A})$, where $\mathcal{X}$ represents the node information and $\mathcal{A}$ represents the edge information [7].

One way to represent graphs with matrices is to use the so-called *adjacency matrix*



(a) Basic graph [10] (b) Types of graph [10] (c) Directed graph [7]

Figure 2-4: Types of graphs. Taken from [10]

$A \in \mathcal{R}^{\mathcal{V}\times\mathcal{V}}$ [12]. The *adjacency matrix* is build so that each node is in a particular row and column of the matrix. The entries of the matrix correspond to the edges between nodes, i.e. if there is an edge, the entry is 1, otherwise it is 0. Formally, $A_{(u,v)} = 1$ if $(u, v) \in \mathcal{E}$, and $A_{(u,v)} = 0$ otherwise. To explain this, it is possible to think in terms of the graph representation of an image, as shown in the figure **2-5**. Note that the figure on the left shows the pixels of a smiling face, the figure on the right shows the graph representation of the image, and the figure in the middle shows the adjacency matrix of the graph. Note that the entries in the adjacency matrix correspond to the nodes that have edges. For example, the node $(0-0)$ is connected to the nodes $(1-0)$, $(1-1)$ and $(0-1)$ [21].

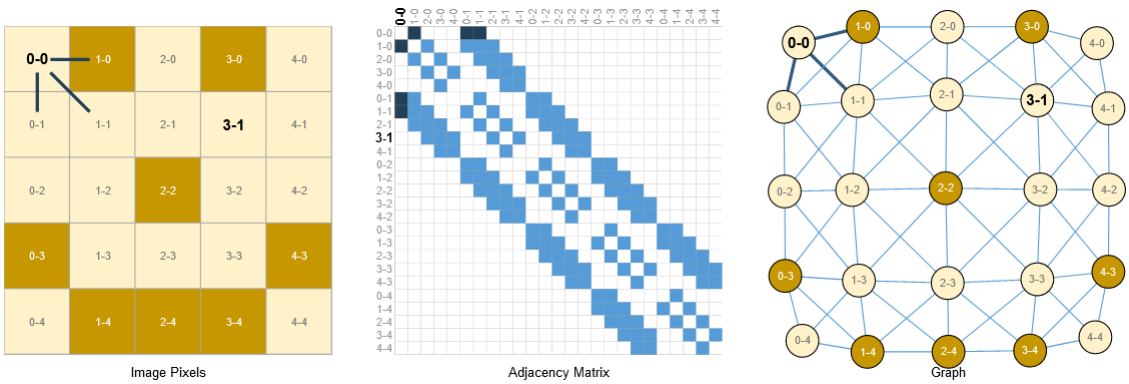


Figure 2-5: Adjacency matrix and graph representation of an image [21].

An alternate method for graph representation involves utilizing a list of edges, often expressed as tuples wherein the first element denotes the edge's source, and the second element designates its destination. This concept is visually conveyed in Figure **2-6**.

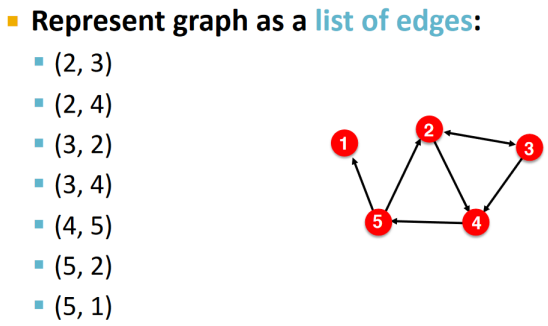


Figure 2-6: Representation of a graph as a list of edges. Taken from [14].

What is graph learning?

Graph learning, also known as machine learning with graphs, involves employing machine learning models to extract attributes from graphs. It's important to recall that a graph comprises various levels of analysis, encompassing nodes and edges. Furthermore, in expansive graphs, sub-graphs can also serve as distinct levels of analysis. These represent the task categories that graph learning encompasses. In essence, the nature of the graph machine learning problems can be categorized based on the level of the task. Consequently, the realms of graph learning categorization encompass node classification (operating at the node level), link prediction (occurring at the edge level), and graph classification (pertaining to the graph level) [10].

Node Classification Node attribute inference involves the prediction of missing or incomplete attributes for nodes based on information from neighboring nodes [10]. In a broader context, node classification seeks to anticipate the label y_u of a given node, which may pertain to a type, category, or attribute [12]. A visual illustration of this scenario can be found in Figure **2-8**, where the left-hand side depicts nodes lacking labels, contrasted with the right-hand side where, courtesy of a classification model, nodes are endowed with labels [21].

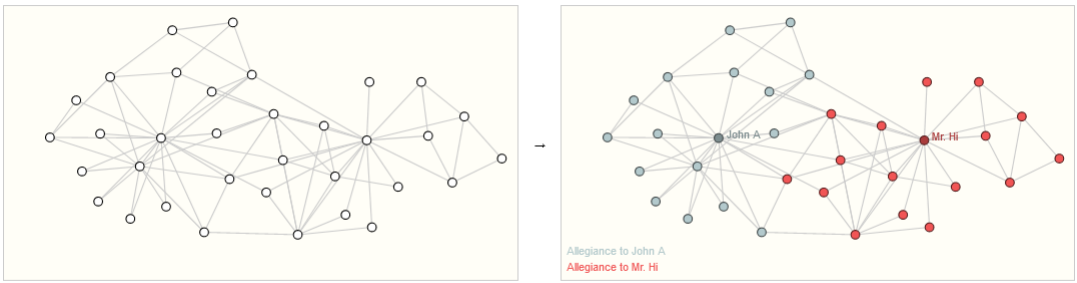


Figure 2-7: Schematic representation for node classification [21]

Link prediction Link prediction pertains to the challenge of deducing absent edges or discovering concealed relationships between entities [10]. This problem arises in cases where explicit connections might be missing due to data collection issues. Moreover, link prediction facilitates prognostications about network evolution consequent to the introduction or removal of new links. This task may manifest under diverse designations, including graph completion or relational inference, contingent upon the specific domain of application [12]. A prime example of link prediction's utility lies in the algorithms powering recommender systems [10]. Figure **2-8** offers an illustrative portrayal of this type of scenario. Given a temporal state at t, link prediction becomes adept at prognosticating relationships in subsequent periods.

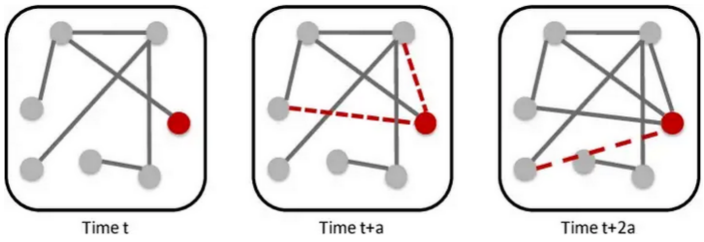


Figure 2-8: Schematic representation of link prediction [16]

Graph Classification This concept involves deducing clusters within a graph, guided by either the graph's inherent structure or the likeness of node attributes [10]. In essence, the objective is to unearth unsupervised patterns of similarity among pairs of graphs [12]. An instance of community detection is exemplified in Figure **2-9**, where the algorithm adeptly identifies ring-shaped structures based on molecular bonds. Once these subgraphs are identified by the algorithm, the entire graph is subsequently categorized.

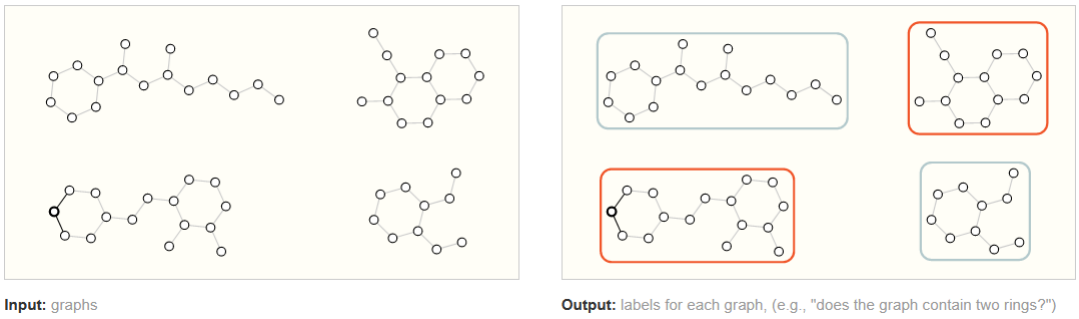


Figure 2-9: Schematic example for community detection. Taken from [21]

Graph Learning Models

According to [24], graph learning models can be categorized into four distinct groups: graph signal processing (GSP)-based methods, matrix factorization-based methods, random walk-based methods, and deep learning-based methods. This classification is visually depicted in Figure **2-10**, with each category further subdivided into sub-methods. Graph signal processing-based methods are concerned with sampling, data recovery, and acquiring the topology structure from data. Random walk-based methods can be segmented into various categories including structure-based random walks, random walks utilizing both structure and node information, random walks within heterogeneous networks, and random walks within time-varying networks. Matrix factorization-based methods can be classified into graph Laplacian matrix factorization and vertex (node) proximity factorization. Lastly, deep learning-based methods encompass a range of techniques such as graph convolutional networks, graph attention networks, graph generative networks, graph spatial-temporal networks, and graph auto-encoder networks.

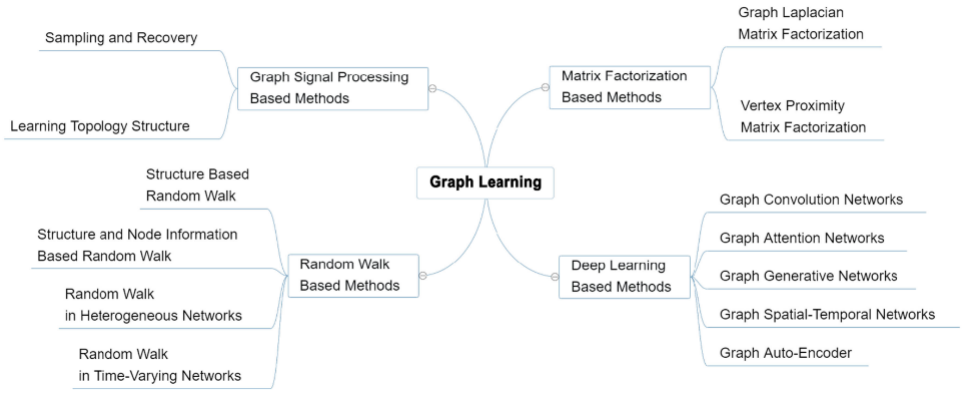


Figure 2-10: Categorisation of graph learning [24]

Deep learning models are designed to generate node representations, often referred to as states. This objective can be accomplished through the utilization of a mechanism known as a 'deep graph network'

Deep Graph Network Also referred to as a graph neural network, the broader framework of deep learning models is illustrated in Figure **2-11**. In essence, the model processes an input graph and subsequently transforms it into node states, all while maintaining the original topology. This implies that each node within the input graph is associated with a corresponding state vector. The models responsible for accomplishing this task are termed deep graph networks (DGNs). It's important to note that DGNs exclusively address the segment of the model responsible for acquiring node representations, without encompassing the predictive aspect.

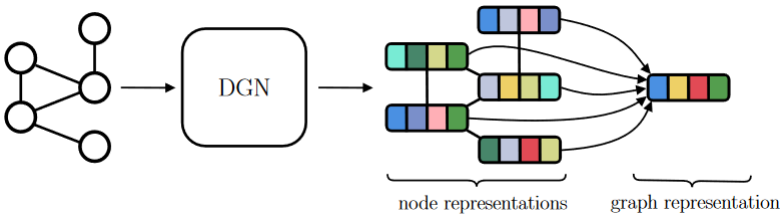


Figure 2-11: Deep graph newtworks schema [7]

As shown in **2-12**, DGNs can be categorized into three distinct groups: Deep Neural Graph Networks, Deep Bayesian Graph Networks, and Deep Generative Graph Networks. The first category encompasses models that draw inspiration from neural architectures. The second category involves the creation of probabilistic models applied to graphs. The third and final category leverages a combination of both neural and probabilistic models to generate graphs [7].

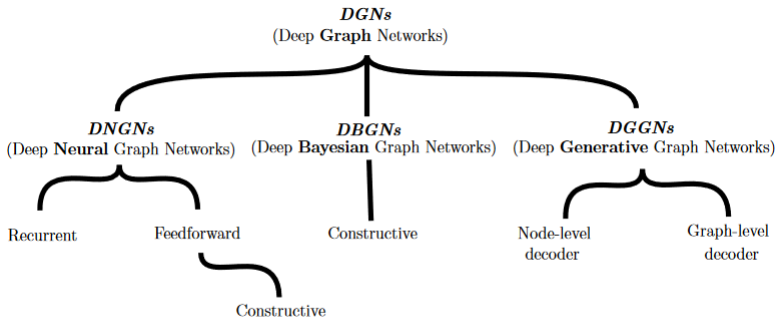


Figure 3: The road-map of the architectures we will discuss in detail.

Figure 2-12: Deep graph newtworks road-map architectures [7]

2.3 Fundamentals of Graph Machine Learning

The strength of graph machine learning lies in its ability to tackle complex structures like graphs, which differ from traditional ML's focus on simple sequences and grids (such as text or images). Graphs possess arbitrary sizes, lack spatial locality, and do not adhere to fixed orders [14]. In this subsection, we will introduce traditional approaches to graph machine learning along with the implementation of graph neural networks (GNNs). The initial part will delve into capturing features across various graph levels (node, link, and graph), as well as node embedding strategies. Subsequently, we will explore the advantages of employing deep learning for node embedding and its application in prediction tasks.

2.3.1 Traditional approaches to graph machine learning

Within graph machine learning, prediction tasks can be executed at different levels: node level, link level, and graph level. To conduct these tasks, as is customary, we require features to train the machine learning model. Hence, the subsequent sections will outline traditional features used for node, link, and graph prediction, with a focus on undirected graphs for simplicity.

Node-level features

At the node level, the primary objective of features is to gather information about the node's position and structure within the network [14]. This objective can be approached through two approximations: importance-based features and structure-based features, each utilizing specific metrics.

Importance-based features This approach captures a node's significance within the graph. Commonly employed metrics include:

Node degree This metric quantifies the number of neighboring nodes. Figure **2-13** illustrates the computation of node degree. Notably, for node A, there's only one neighbor, whereas node D has four neighbors. Consequently, node A's degree (k_A) equals 1, while node D's degree (k_D) is 4. A limitation of this metric lies in its inability to weigh neighbor importance. To address this, node centrality metrics are used, which consider the nodes' significance within the graph.

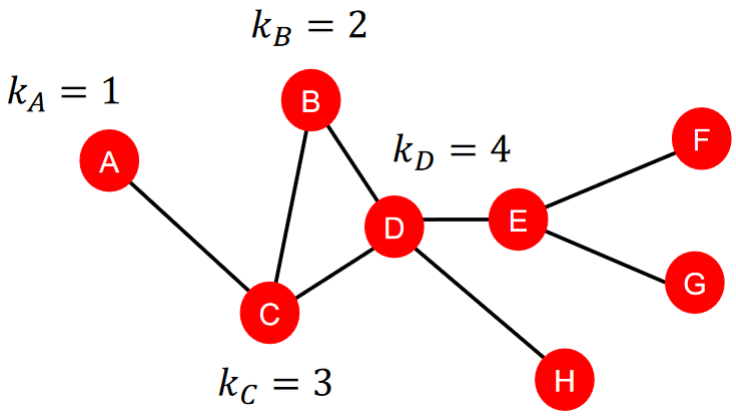


Figure 2-13: Schematic representation of a graph and node degree computation. Taken from [14]

Node Centrality This metric quantifies the significance of neighboring nodes within a graph. It can be calculated using various options, such as eigenvector centrality, betweenness centrality, closeness centrality, among others.

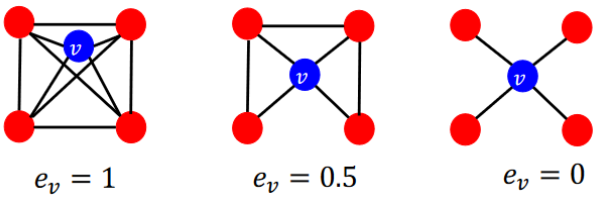
Structure-based features This approach captures the topological properties of the local neighbourhood around a node [14]. The common metrics are

Node Degree Counts the number of neighbouring nodes as mentioned above.

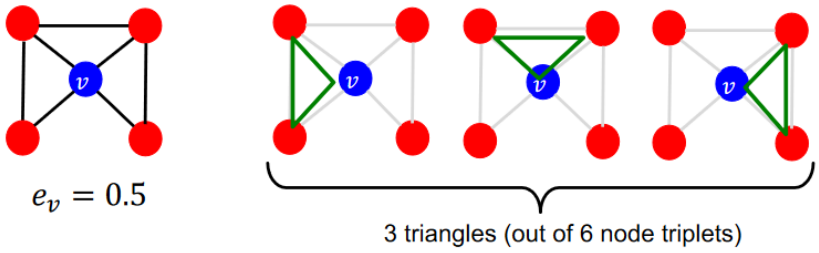
Clustering coefficient Measures how well neighbouring nodes are connected. Mathematically defined as

$$c_v = \frac{|(v_1, v_2) \in \mathcal{E} : v_1, v_2 \in \mathcal{N}(u)|}{\binom{k_v}{2}}, \tag{2-1}$$

where the numerator counts the number of edges between neighbours of node $\mathcal{N}(u)$ and the denominator computes the number of pairs of nodes in neighbourhood u [12]. The clustering coefficient can be interpreted as counting the number of triangles in the network. The way to generalize this counting of subgraphs can be achieved by using graphlets. Figure **2-14** shows an example of how to compute the clustering coefficient.



(a) Clustering coefficient compute from mathematical formulation.



(b) Clustering coefficient understood as a count of triangles.

Figure 2-14: Schematic representation of a graph and clustering coefficient computation. Both images taken from [14]

Link-level features

The objective of link-level features is to capture attributes for pairs of nodes. The associated task involves predicting new connections based on the existing connections within

the graph [14]. There are two formulations for the link prediction task: one for predicting missing links that are randomly absent, and another for cases where links change over time. In general, the methodology involves selecting a node pair (x, y) and computing a score $c(x, y)$, then retaining the top n pairs as new connections. Link-level features can be derived through three approaches: distance-based features, local neighborhood overlap, and global neighborhood overlap.

Graph-level features

The method for capturing graph-level features involves utilizing graph kernels, which are employed to assess the similarity between two graphs. This can be accomplished using various types of kernels, with the most frequently used ones being graphlet kernels and Weisfeiler-Lehman kernels [14].

Node embedding

An important phase in graph machine learning is node embedding, which involves transforming graph elements into an embedding space. This process aims to encode intricate graph information into simplified representations that can be employed for predictive tasks [14]. According to [12], this objective can be accomplished through an encoder-decoder perspective. In the encoder stage, each node in the graph is mapped into a low-dimensional vector. Subsequently, the decoder phase takes the low-dimensional node embedding and employs it to reconstruct information about the node in the original graph. The encoder-decoder framework is illustrated in Figure **2-15**.

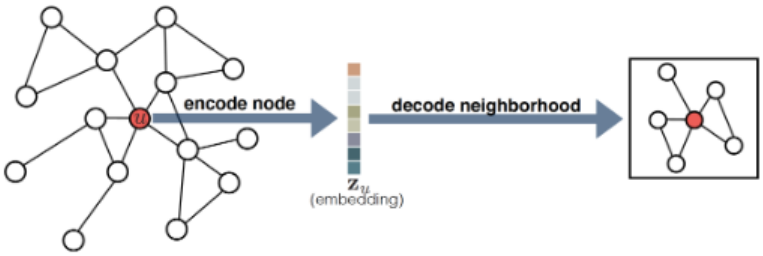


Figure 2-15: Schematic representation of node encode-decode framework. Taken from [12]

Formally, the encoders are functions that maps the nodes $v \in \mathcal{V}$ to vector embeddings $z_u \in R^d$. An illustration of the encoder functions is provided in Figure **2-16**.

The simple structure for the encoder has the form, [12]

$$ENC : \mathcal{V} \longrightarrow \mathcal{R}^d. \tag{2-2}$$

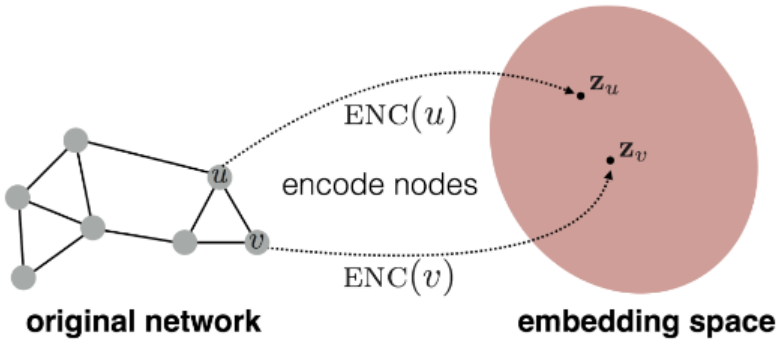


Figure 2-16: Schematic representation of node embedding problem. Taken from [12]

The objective of the encoder is to map nodes in such a way that the similarity in the embedding space closely mirrors the similarity in the original graph [14]. This is achieved as follows:

$$Similarity(u, v) \approx Similarity(z_u, z_v), \tag{2-3}$$

where $Similarity(u, v)$ corresponds to the similarity of u and v in the original graph and, $Similarity(z_u, z_v)$ corresponds to the similarity of the mapped z_u and z_v nodes in the embedding space. To simplify the notation, some authors use S_G to refer to similarity in the original graph and S_E to refer to similarity in the embedding space.

According to [12], the most common embedding approach is known as *shallow embdeing* and consists of implementing the encoder as a lookup matrix, such as

$$ENC(v) = \boldsymbol{Z}_v = \boldsymbol{Z} \cdot \boldsymbol{v}, \tag{2-4}$$

where $\boldsymbol{Z}$ is the embedding matrix, in which each node is a column, and $\boldsymbol{v}$ is the indicator vector that allows to identify the node v. In this way, each node is associated with a unique embedding vector.

Apart from *shallow encoding*, there exist other encoding architectures that extend the utilization of node features or local graph structure, often referred to as graph neural networks [12]. Selecting the appropriate encoder is a crucial step, and it's equally important to choose the similarity functions accurately. The forthcoming sections will delve into shallow encoding options that employ random walks as a technique to create the lookup embedding matrix, as seen in the cases of deep-walk and node2vec.

Deep walk embedding This embedding technique consists in defining the similarity function between nodes S_G and the similarity function in the embedding space S_E. The similarity

between vertices is defined as the probability of visiting the vertex u by a random walk on the graph, starting from the vertex v. This is

$$S_G(u, v) = P(u|v). \tag{2-5}$$

Similarity in embedding space will defined by softmax function and the product of the embedding nodes, as follows:

$$S_E(z_u, z_v) = \frac{\exp(z_u^T z_v)}{\sum_{n \in V} \exp(z_u^T z_v)}. \tag{2-6}$$

Hence, starting with random values for z_u and z_v, is possible to find the correct embedding space such that a loss function is minimised. For this case the loss functions is defined as

$$\mathcal{L} = \sum_{(u,v) \in D} -\log(S_E(z_u, z_v)) \tag{2-7}$$

Node2vec Embeddings The primary distinction between deep-walk embedding and node2vec embedding lies in the way S_G is defined. Deep-walk employs an unbiased random walk, signifying that each path carries the same weight. Conversely, node2vec involves a biased walk, resulting in paths with varying weights. As a consequence, while S_G retains the same mathematical structure, the probabilities are influenced by bias [12].

To enrich the randomly generated paths with weights, node2vec employs two search strategies: breadth-first search (BFS) and depth-first search (DFS). Figure **2-17** illustrates the difference in exploration between these methods. These biased random walks offer insight into both the local and global characteristics of the path [14].

2.3.2 Graph neural networks

Graph Neural Networks (GNNs), also known as Deep Graph Networks, serve as tools for node embedding. In the previous subsection, we introduced an encoding technique referred to as "shallow encoding," which essentially involves mapping nodes into a lookup matrix. However, as highlighted in [14], this encoding method has certain limitations: its complexity scales as $O(|v|)$, resulting in linear growth relative to the number of nodes; each node possesses its own embedding without parameter sharing; it lacks transductive capability, making it unable to generate embeddings for new, unseen nodes; and it does not incorporate node features.

In contrast, GNNs offer an encoding strategy that leverages neural networks to devise coding functions. As elucidated in subsection 2.2, GNNs (or Deep Graph Networks) facilitate the

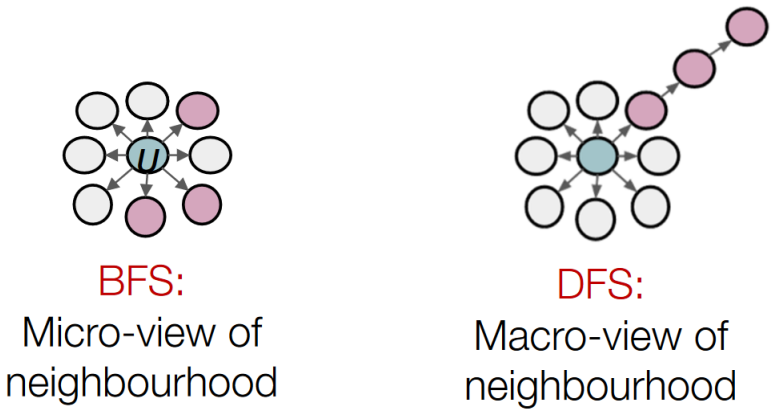


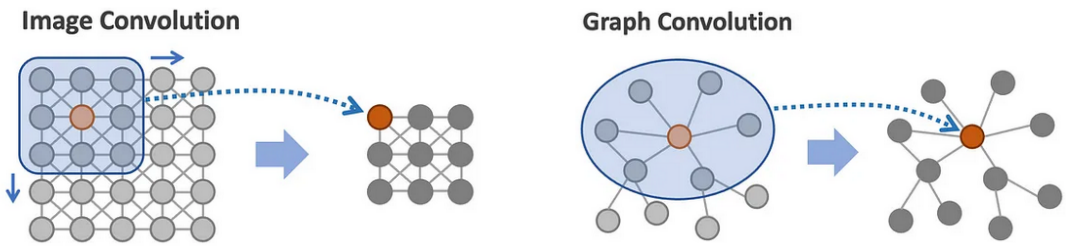
Figure 2-17: Schematic representation of bread first search and deep first search. Taken from [14]

translation of graph structures into vector representations. Post-embedding, various tasks such as regression or classification can be executed through standard ML algorithms. GNNs are versatile and capable of tackling node classification, link prediction, community detection, and network similarity tasks [14].

The subsequent sections will delve into the GNN framework, exploring prominent GNN algorithms like Graph Convolutional Networks (GCN), GraphSAGE, and Graph Attention Networks (GAT). Lastly, we will explore the application of GNNs to prediction tasks.

GNN Framework Graph Convolutional Networks (GCNs) can be comprehended by drawing parallels with Convolutional Neural Networks (CNNs). While conventional neural networks process vectors, convolutional networks can accommodate data in the form of matrices or tensors. This versatility is achieved through the utilization of a kernel to convolve the original matrix with another. The convolution process is illustrated in Figure **2-18**a, demonstrating how a group of pixels is amalgamated into one through the convolution, depicted by the blue arrow. Kernels come in diverse types, and their selection stands as a pivotal facet in deep learning algorithms.

Similarly, as depicted in Figure **2-18**b, the graph convolution process aims to transform one node into another. However, in this context, the underlying graph structure remains unchanged while information about the nodes evolves. This process of graph convolution can be succinctly summarized, as shown in Figure **2-19**, where a target node and its neighbors are visualized as a network. The embedding outcome is contingent on both the conveyed message and the aggregation method employed. Consequently, graph neural networks can be perceived as diverse methodologies that delineate messaging and aggregation functions to facilitate node embedding.



(a) Scheme of image convolution. (b) Scheme of graph convolution.

Figure 2-18: Schematic representation of convolution process in images and graphs. Both images taken from [15]

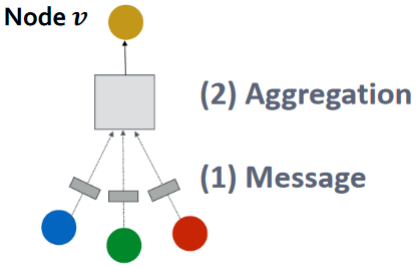


Figure 2-19: Schematic representation of the GNN framework. Taken from [14]

GCN The main idea is to transform the information of a node's neighbours and combine them to update the node's information. This idea is realised by the definition of a computation graph. An example is shown in **2-20**, where given a simple graph on the left, if the target node A is the target, its neighbours define the computation graph on the right.
In the same figure, in the scheme for the computation graph on the right, there are boxes in front of each node. These represent the way in which the information from the incoming neighbours is aggregated. Note also that the computation can be of arbitrary depth, nodes have embeddings at each layer of the computation graph in such a way that for layer 0 embeddings of node u it is input feature x_u and at a layer k embeddings gets information from nodes that are k hops away [14]. The boxes in the figure **2-20** represent the approaches to aggregating information across layers. There are two ways of aggregating information, by calculating the average messages from neighbours or by using a neural network. Mathematically, the updating can be written as

$$h_v^{l+1} = \sigma\left(W_l \sum_{u \in \mathcal{N}(v)} \frac{h_u^{(l)}}{|\mathcal{N}(v)|} + B_l h_v^l\right), \forall l \in \{0, 1, ..., L-1\}, \tag{2-8}$$

where h_v^l represents the embedding after L layers of neighbourhood aggregation, $\sigma(x)$ is the nonlinear activation function, $\sum_{u \in \mathcal{N}(v)} \frac{h_u^{(l)}}{|\mathcal{N}(v)|}$ corresponds to the average of the previous layer embedding of the neighbour, and L is the total number of layers.

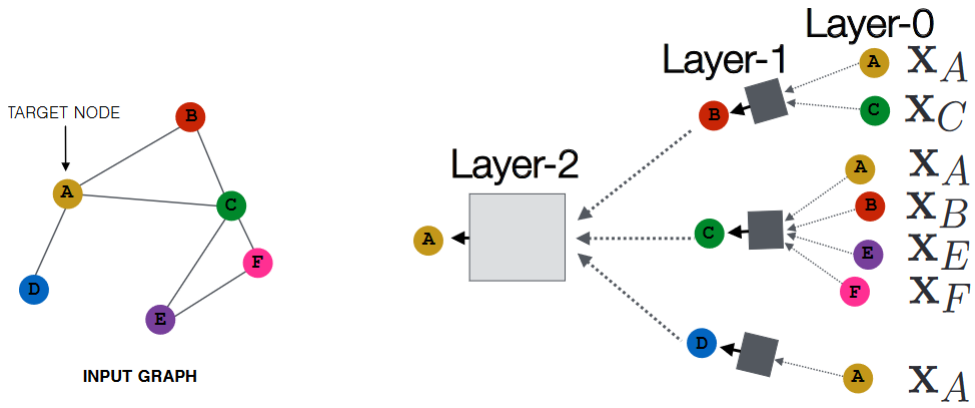


Figure 2-20: Schematic representation a computation graph. Taken from [14]

Due to the fact that all nodes have the same aggregation parameters, the model can be generalized to unseen new nodes or new graphs. This means that GNN are inductive models [14].

GraphSAGE This algorithm constitutes a variation of Graph Convolutional Networks (GCN), differing primarily in the approach to information aggregation. This modification introduces a more adaptable aggregation strategy, allowing the equation 2-8 to be extended as presented in [14].

$$h_v^{(l+1)} = \sigma\left(W_l \cdot AGG(h_u^{(l)}, \forall u \in \mathcal{N}(v), B_l h_v^{(l)})\right), \tag{2-9}$$

in particular, if AGG function corresponds to the weighted average of neighbours it result into the GCN presented bellow.

GAT This algorithm add a weighting factor on the aggregation function, this is,

$$h_v^(l) = \sigma\left(\sum_{u \in \mathcal{N}(v)} \alpha_{vu} W^{(l)} h_u^{(l-1)}\right), \tag{2-10}$$

where $\alpha_{vu} = \frac{1}{|\mathcal{N}(v)|}$ is the weighting factor of the message of node u to node v. This is the attention factor, and allows the graph to focuses on the important information of the input data due that not all nodes are equally important [14].

Conclusions

This chapter serves as an introduction to various machine learning methods, the characterization of graph element statistics, and the fundamentals of graph-based machine learning. It underscores the distinctions between artificial intelligence and machine learning, with a specific focus on categorizing machine learning into classical and deep learning paradigms. Furthermore, the handling of structured and unstructured data is addressed, along with the limitations of conventional machine learning techniques when confronted with complex data structures.

The chapter also introduces the concept of graph learning and its applicability in addressing challenges related to graph-structured data. It presents a mathematical framework for interacting with graphs, outlines strategies for implementing node embedding, and explores common algorithms employed in applying machine learning across node, link, and graph levels. The overarching goal of this chapter is to provide a comprehensive grasp of the foundational principles of graph machine learning and its potential for tackling intricate problems.

3 ALIANZA EFI CO-AUTHORSHIP NETWORK ANALYSIS

In this chapter, the results obtained from the implementation of graph machine learning techniques on the co-authorship network will be presented. Specifically, three tasks were performed: node prediction, link prediction, and community detection. Node prediction was carried out using MLP, GCN, GraphSAGE, and GAT. Link prediction was accomplished using the encoder-decoder approach. Community detection was achieved through GCN embedding and t-SNE dimensionality reduction.

3.1 Project Alianza EFI

Before describing Alianza EFI's project, it is necessary to provide an overall context of this research platform. To do so, it is essential to mention the program Colombia Científica, a government initiative aimed at incentivizing scientific projects to enhance the quality of higher education in Colombia. This program supports research and innovation to foster the development of various regions within Colombia. Additionally, it serves to identify problems and solutions for the productive sector. Colombia Científica comprises multiple institutions that form alliances to contribute to the project, with Alianza EFI being one of these alliances [4].

Alianza EFI was selected as one of the alliances in response to the Colombia Científica project call. They presented a scientific program titled "Productive and Social Inclusion: Programs and Policies for the Promotion of a Formal Economy." This proposal received government funding and is planned to be executed over a span of four years, starting from the year 2019. The role of Universidad del Rosario in this alliance is crucial, as it is the lead institution responsible for managing the allocated resources [4]. The overarching objective of Alianza EFI is to diagnose, analyze, and address the factors and barriers that impact the productive and social inclusion of economic agents in diverse contexts, all from a systemic perspective [4].

In total, Alianza EFI consists of eighteen participating institutions. This includes four international partners, seven national higher education institutions, and seven national entities from the productive sector [4]. Those institutions are:

- **Governance** Universidad del Rosario,

- **Higher education institutions** Universidad de Antioquia, Universidad del Valle, Universidad del Quindio, Universidad autónoma latinoamericana, Universidad Minuto de Dios, Universidad de Ibagué;
- **Articulatory entities** Federación Nacional de cafeteros de Colombia, Asociación nacional de cajas de compensación familiar, Asobancaria, Cámara Colombiana de la construcción;
- **Companies** Fundación Avina, Fundación capital, Asociación de mujeres afrodescendientes del norte del cauca;
- **International allies** University of illinois at chicago, University of Oxford, University of Milano-Bicoca and, University of Pennsylvania.

Similarly, Alianza EFI is organized into eight projects, each addressing different dimensions of informality. These projects are as follows:

1. Entrepreneurship, Development of Business Capacities and Productive Inclusion,
2. Inclusive work Markets,
3. Rural Economic Informality,
4. Cities as Scenarios for Social Inclusion,
5. Macroeconomic and Institutional Aspects on the Causes and Consequences of Informality,
6. The Mind of the Informal Economic Agent: Preferences, Abilities and Social Norms,
7. Social lab, and,
8. Institutional strengthening.

Hence, the various papers produced by the alliance were created under the framework of these projects. For the period spanning from 2019 to 2021, a total of 43 participating institutions and 178 contributing authors were involved. This collaboration resulted in the production of 122 scholarly outputs. The individual contribution of each project is detailed in Table **3-1**.

Furthermore, it is evident that the alliance has fostered a large and diverse team of authors and institutions. Beyond the scientific output, one intangible yet crucial aim of the alliance is to foster the growth of a robust scientific ecosystem that strengthens relationships among its authors and institutions. With this objective in mind, by visualizing the relationships between authors and institutions as a network, it becomes possible to identify key nodes that

play significant roles within the network. This could lead to the identification of potential author contributions, inter-author relationships, and thematic communities.
Hence, the contribution of this work to the Alianza EFI project involves comprehending the construction of the ecosystem and scientific agenda surrounding the products developed by the project's contributors. In light of this objective, three hypotheses are formulated:

Identification of Key Contributors By employing node prediction techniques on the co-authorship network, it will be possible to identify pivotal authors who play significant roles within the Alianza EFI project. These authors contribute substantially to research topics, product development, and meaningful collaborations with other authors. The identification of such key authors will assist Alianza EFI in resource allocation, expertise recognition, and the facilitation of new fruitful collaborations.

Prediction of New Collaborations The application of link prediction techniques on the co-authorship network will facilitate the identification of potential novel collaborations among authors. The establishment of these new connections will empower Alianza EFI to cultivate fresh internal and external institutional partnerships, thereby enhancing its research capacity.

Detection of Topic Communities Utilizing community detection techniques on the co-authorship network will enable the discovery of groups of authors who share common research interests. These communities can offer complementary perspectives to the diverse projects undertaken by Alianza EFI.

Project	**Products**
Entrepreneurship, Development of Business Capacities and Productive Inclusion	24
Inclusive work Markets	31
Rural Economic Informality	8
Cities as Scenarios for Social Inclusion	17
Macroeconomic and Institutional Aspects on the Causes and Consequences of Informality	21
The Mind of the Informal Economic Agent: Preferences, Abilities and Social Norms	14
Social lab	7

Table 3-1: Developed products per Alienza EFI's project.

3.2 Methods

This section describes the hardware and software used for the development of the project. Also, provides a description of the acquisition, preparation and exploration of the working data.

3.2.1 Hardware

The hardware used in this project correspond to a machine with the following specifications:

Processor 11th Gen Intel(R) Core(TM) i5-11300H @ 3.10GHz
RAM 24GB
SSD 512GB
GPU NVIDIA GeForce RTX 3050

3.2.2 Software

Graph analysis can be conducted using various programming languages and frameworks. However, due to Python's growing popularity in the field of machine learning, driven by its performance, robust community, and extensive documentation, this project will be implemented using Python. Python's ecosystem offers numerous frameworks for deep learning, including NetworkX, PyTorch Geometric, Deep Graph Library (DGL), Graph Nets, Spektral [2], Apache spark and neo4j [16], and, StellarGraph [10]. For this project, the selected frameworks are NetworkX and PyTorch Geometric. NetworkX will be used for creating graph structures, while PyTorch Geometric will be employed to perform computations involving deep learning networks.

PyTorch Geometric PyTorch Geometric is a Python framework designed to facilitate working with irregular data structures and graph neural networks [3]. It provides a seamless integration with PyTorch, enabling the use of dataset and dataloader classes tailored for graph data. Additionally, PyTorch Geometric supports GPU acceleration to enhance computational efficiency for deep learning algorithms. Installation can be carried out through package managers like PIP, Conda, or Mamba. According to Neptune AI [2], PyTorch Geometric is the most popular graph repository with 11.2K stars, followed by DGL with 7.4K stars and Graph Nets with 4.9K stars. Therefore, PyTorch Geometric is the chosen framework for this work.

NetworkX NetworkX is a Python package that provides tools for analyzing and studying graph data structures [6]. Widely used within the academic community, NetworkX finds applications in diverse domains ranging from biological structures to transportation systems. The advantages of NetworkX include its straightforward implementation, built-in functions for computing graph statistics such as centrality coefficients and clustering coefficients, and support for both directed and undirected graphs. NetworkX can be installed using PIP or Conda.

3.2.3 Data acquisition, pre-processing and exploration

As is usual in machine learning tasks, the initial steps involve data acquisition, pre-processing, and exploration. This subsection outlines the data acquisition process, the preparation of data to transform it from a structural format to a graph representation suitable for NetworkX. Additionally, it provides a descriptive analysis of the institutions, authors, and papers involved.

Data acquisition

The data required for applying graph machine learning algorithms to analyze the co-authorship network consists of information about authors' products, affiliation institutions, and produced papers. This information was obtained from a database managed by Alianza EFI. The database contains details such as paper titles, author names, affiliated institutions, and EFI-related projects. However, abstract information was absent from this database, and abstracts are crucial inputs for training the algorithms. To collect abstracts, a web scraping script was developed to extract this information from the EFI website. The subsequent sections provide a breakdown of the available information.

Alianza EFI products database The database comprises 18 fields and 665 records, with each record corresponding to an author and their produced products. Specifically, the table contains information about papers produced from January 2019 to February 2023. The dictionary structure is presented in Table **3-2**.

Field name	Data type	Description
Project associated with the product	String	EFI project associated
Intellectual Property Area	String	not used
Id	String	Paper unique identifier
Type of product to be reported	String	not used
NAME/PRODUCT TITLE	String	Title of the paper
Product progress level in %	String	not used
Entities directly related to the product	String	not used
Alliance product authors	String	Name of the Alianza EFI authors
Institution	String	Institution of the Alianza EFI authors
Other Product Authors	String	Name of the external authors
Institution	String	Institution of the external authors
Research groups associated with the product	String	not used
Percentage of participation of each author and institution in the product	String	not used
Type of support or annex that accompanies the reported product	String	not used
Comments and/or observations about the product	String	not used
Product report date	String	not used
Link	String	not used

Table 3-2: Data dictionary for EFI products database. The table contains 665 registers.

Papers abstracts After applying web scrapping to the Alianza EFI web site, it was possible to capture information for 255 papers. This information was complemented by the Alianza EFI team, with additional 90 abstracts for papers that where not published yet in the EFI web site, but where related in the EFI products database. Hence, there are a the total amount of 345 abstracts. The resulting table is summarise with the data dictionary shown in Table **3-3**.

Field name	**Data type**	**Description**
Title of the product	String	Title of the paper
Abstract	String	Abstract of the paper
Authors	String	Authors of the paper

Table 3-3: Data dictionary for EFI products abstracts. The table contains 274 registers.

Data pre-processing

The source data was derived from two distinct tables: the Alianza EFI products database and the Papers abstracts table. It is important to emphasize that both of these tables consist of structured data, which needs to be transformed from a tabular format into a graph representation. As discussed in Chapter 2, a graph is comprised of nodes and edges. In the context of co-authorship analysis, each node represents an author, while edges connecting nodes represent collaborations between authors. Consequently, constructing the co-authorship network requires defining the connections between authors (edges) and attributes associated with each author (nodes).

Authors links The strategy to represent authors links is through the use of list of edges representation. This means that the graph will be explicitly expressed as a list of tuple of authors whom collaborated in at least one paper (i.e. [(author 1, author 2), (author 1, author 3), ..., (author n, author m)]). However, we take into account also those authors that do not show collaborations (i.e. [(author 4,), (author 5,), ..., (author p,)]). Hence, identifying unique authors and its collaborations, the resulting graph can be summarised as follows:

- Number of nodes: 390,
- Number of edges: 318,
- Has isolated nodes: True,
- Has self-loops: False,
- Is undirected: True.

Authors attributes Once the graph is built, it becomes necessary to assign attributes to the nodes, encompassing features and labels for each author. The label assigned to each author corresponds to the EFI project with which the author is most significantly associated. On the other hand, author attributes were computed using Natural Language Processing (NLP) techniques applied to the abstracts of the papers the author has collaborated on. This process involved creating a corpus and subsequently vectorizing that corpus on a per-author basis.

For the corpus creation, the approach employed was to identify the abstracts of the various papers in which each author participated and then concatenate these abstracts to generate a unified corpus for each author. Ultimately, the comprehensive corpus was composed of the amalgamation of all Alianza EFI authors' individual corpora. Figure **3-1** provides a graphical illustration of how the full corpus was constructed.

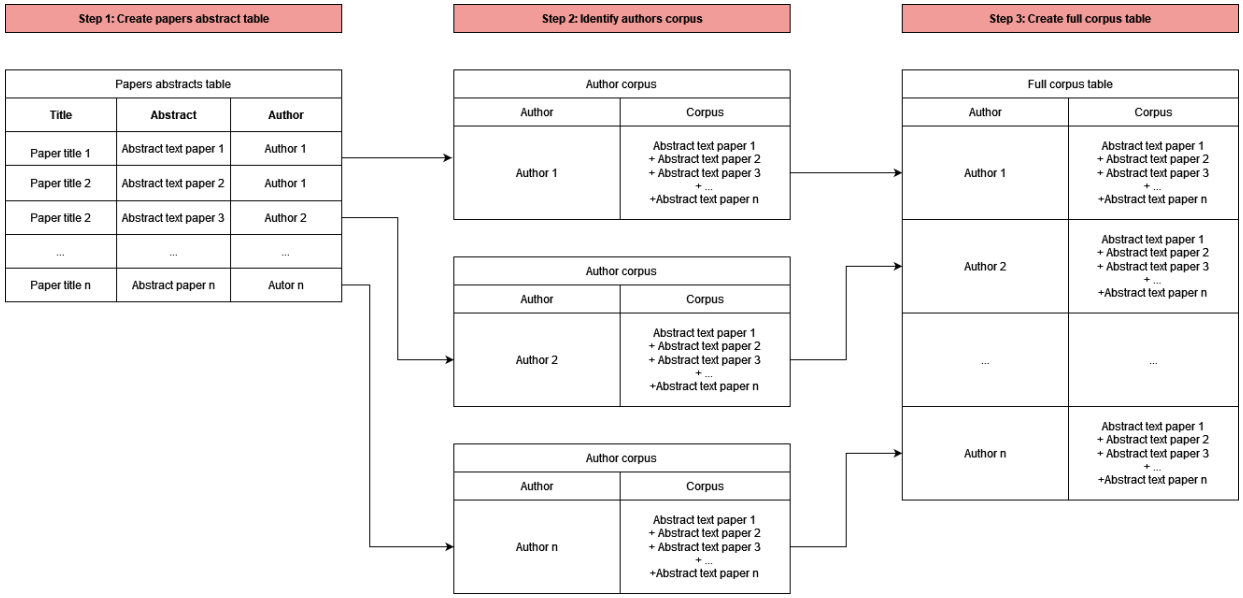


Figure 3-1: Process to build corpus of abstracts per author. From the papers abstracts table, identify abstracts for each author to form a unique corpus, later, the full corpus table will be the corpus of all the authors.

Having established the corpus table, the technique utilized for vectorizing each author involved the implementation of the Term Frequency Inverse Document Frequency algorithm (TF-IDF). This widely employed algorithm is adept at transforming textual content into numerical representations [9]. Notably, TF-IDF is well-suited for this project, as it permits the quantification of a word's significance within a localized context (an author's corpus) and the subsequent comparison of this significance against a broader context (the full corpus). Following the application of the TF-IDF algorithm to the comprehensive corpus, and accounting for words appearing with a minimum frequency of 10% and a maximum frequency

of 95%, a total of 98 salient words were identified. Consequently, the vectorization process endowed each author with 98 features. The ten most frequently occurring words across the entire corpus were found to be "increase," "model," "informality," "find," "use," "work," "Colombia," "study," "tax," "labor market," and "information." Additionally, to harness the informative potential of the graph structure, node statistics were appended as attributes to each author. This entailed augmenting the 98 features from the previous outcome with centrality coefficient, clustering coefficient, and node degree metrics. Consequently, each author possessed 101 features. With the edges, nodes, node labels, and attributes duly defined, it became possible to visualize the co-authorship network, as depicted in Figure **3-7**.

Finally, an overall scheme of the process followed for the development of this project is presented in figure **3-2**. The differents steps followed were:

Source Data This phase involves obtaining the raw data, including the original EFI products database, utilizing web scraping techniques to retrieve abstract information from the EFI website, and supplementing information provided by the Alianza EFI team.

Preprocessing This phase encompasses the exploration and cleansing of the raw data, ensuring its quality and suitability for subsequent analysis.

Data Transformation In this phase, the necessary inputs are generated to construct graphs in Python, forming the foundation for further analysis.

Graph Structure This phase focuses on creating the graph structures using NetworkX and PyTorch within the Python environment.

Machine Learning Techniques During this phase, various machine learning techniques are applied at the node, link, and community levels to address the formulated hypotheses.

Results In this phase, the outcomes obtained from the application of each machine learning technique are presented and analyzed.

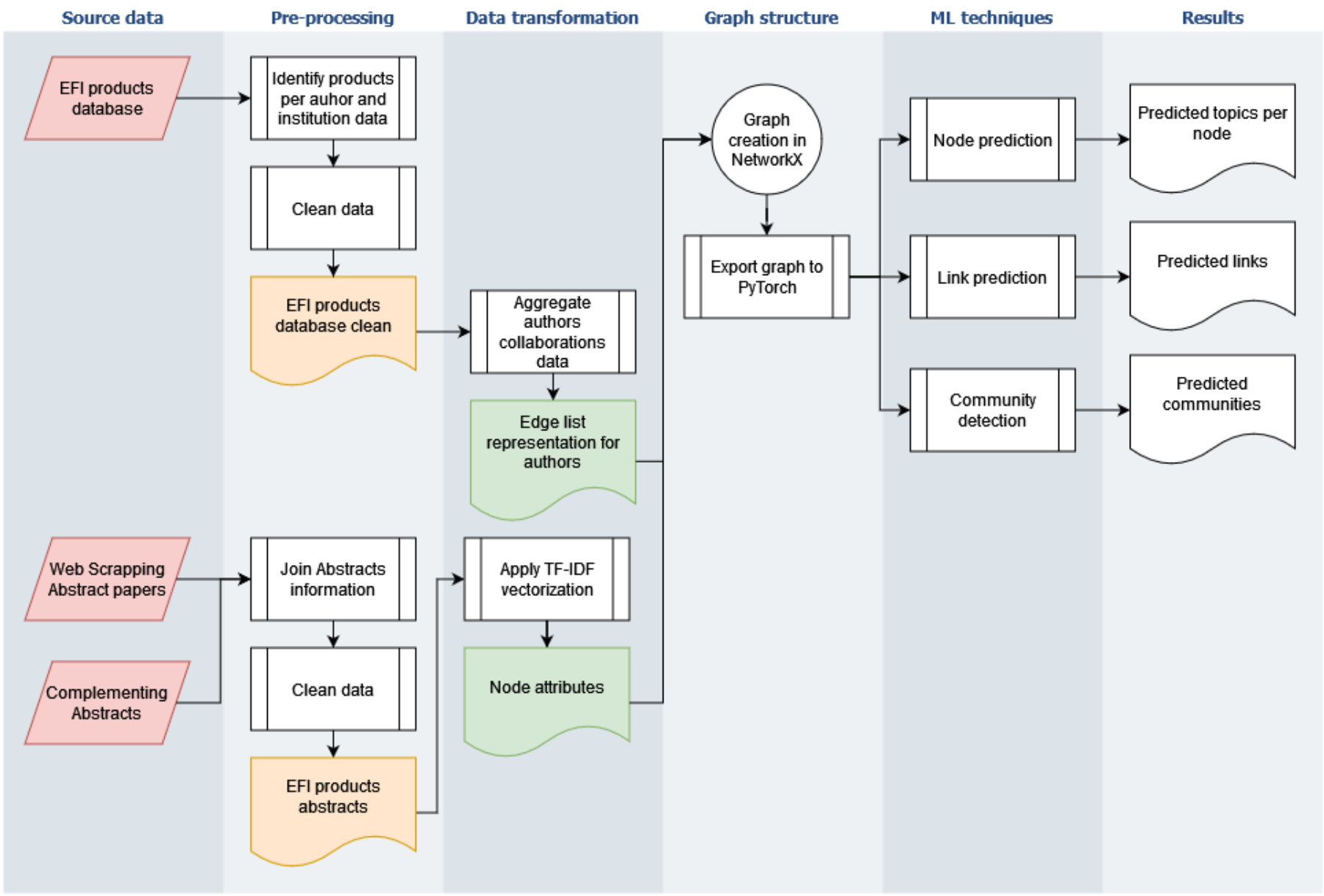


Figure 3-2: Schemetic representation for the process implemented for the development of this project.

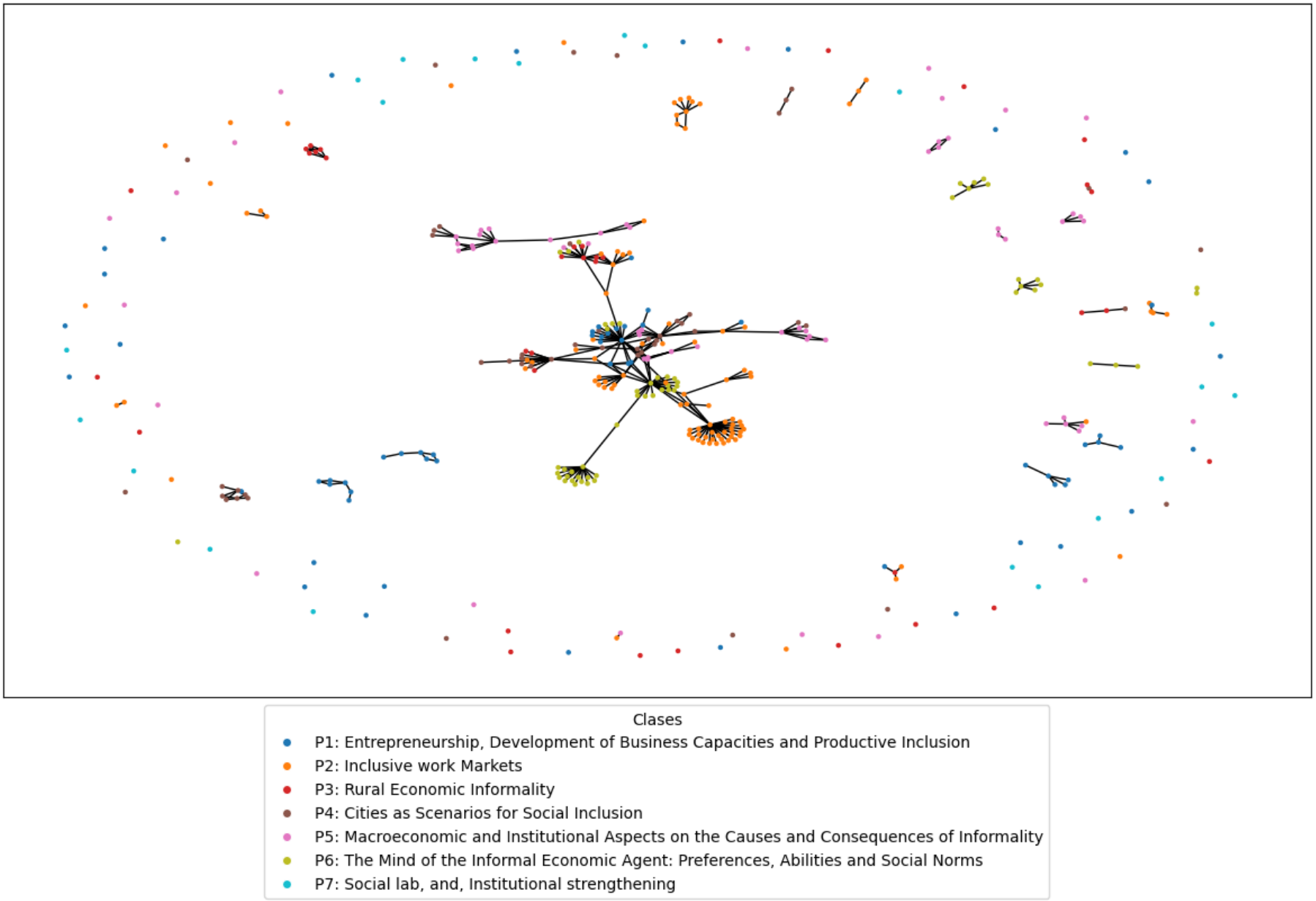


Figure 3-3: Graph representation of the co-authorship network by author of the Alianza EFI project authors.

Exploratory data analysis

From the descriptive perspective, it is possible to conclude that the Alianza EFI has garnered contributions from 390 distinct authors, affiliated with 112 unique institutions, resulting in the creation of 274 distinct products. Subsequently, this section will offer a comprehensive overview of the Alianza EFI projects, followed by an examination from the viewpoint of individual authors and institutions. Throughout the following analysis, please consider the following notation:

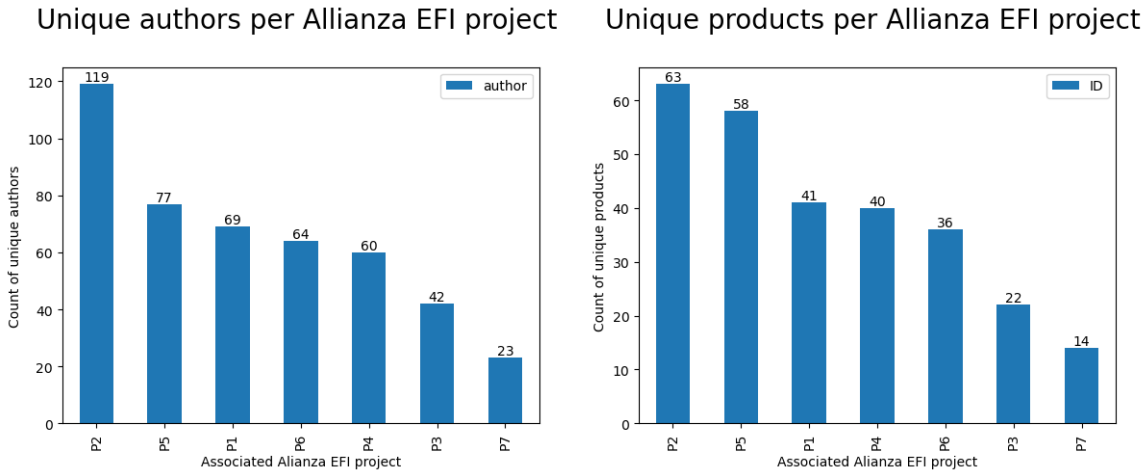
- P1: Entrepreneurship, Development of Business Capacities and Productive Inclusion,
- P2: Inclusive work Markets,
- P3: Rural Economic Informality,
- P4: Cities as Scenarios for Social Inclusion,
- P5: Macroeconomic and Institutional Aspects on the Causes and Consequences of Informality,
- P6: The Mind of the Informal Economic Agent: Preferences, Abilities and Social Norms,
- P7: Social lab, and, Institutional strengthening.

Descriptive analysis per Alianza EFI project Figure **3-5** illustrates the performance of each Alianza EFI project in terms of the number of authors and products associated with them. Notably, the projects that stand out are P2, boasting the involvement of 119 authors and the creation of 63 products; P5, which includes 77 authors and yields 58 products; and P1, encompassing 69 authors and resulting in 41 products.

Descriptive analysis per Alianza EFI authors Figure **3-5**(a) presents the top ten most productive authors within Alianza EFI. The leading three authors are Cesar Mantilla, contributing to 26 products; Paul Rodriguez, also contributing to 26 products; and Juan Miguel Gallego, with 20 products.

On the other hand, for identifying those who collaborate most extensively with other authors, refer to Figure **3-5**(b), which displays authors with more than 10 collaborations. Notably, the three authors with the highest number of collaborations are Darwin Cortes, engaged in 33 collaborations; Paul Rodriguez, collaborating in 27 instances; and Cesar Mantilla, collaborating in 22 instances.

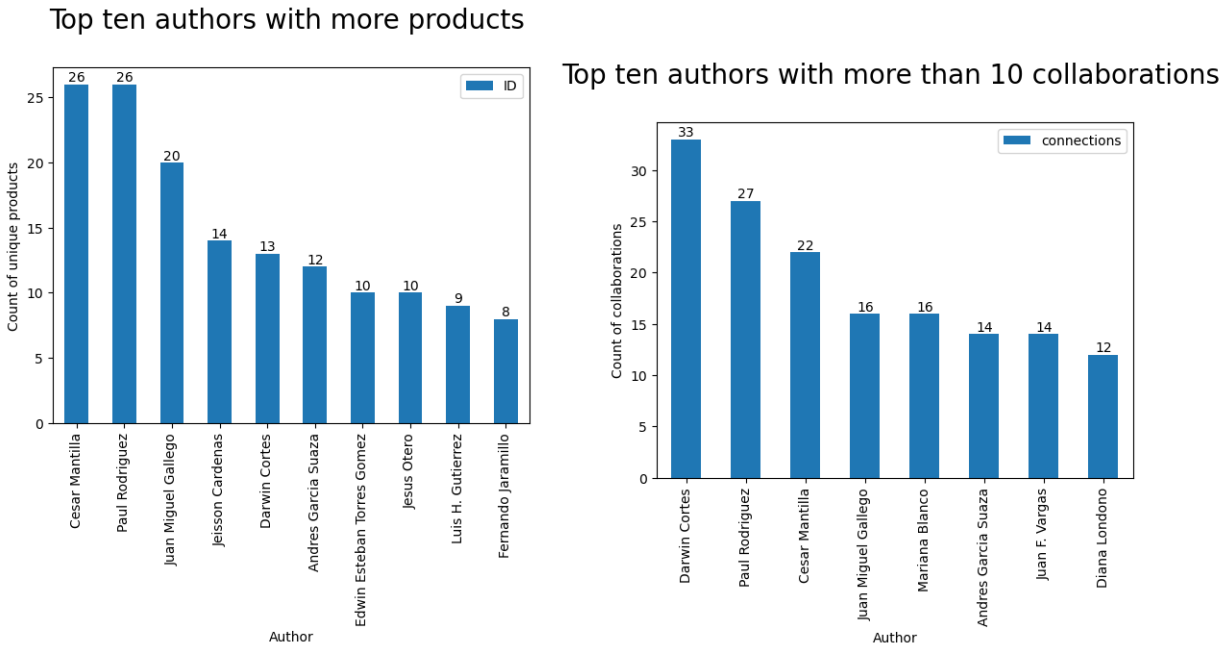
When examining the co-authorship network depicted in Figure **3-7**, it becomes evident that numerous authors exist in isolation. Consequently, this outcome highlights an opportunity to foster collaborations among diverse authors from various Alianza EFI projects.



(a) Count of unique authors per Alianza EFI project.

(b) Count of unique products per Alianza EFI project.

Figure 3-4: Count of authors and products per Alianza EFI project.



(a) Count of unique products per author.

(b) Count of unique collaborations per author

Figure 3-5: Count of products and collaborations per author of the Alianza EFI project.

Descriptive analysis per Alianza EFI institutions The graphical representation of inter-institutional relationships is provided in Figure **3-7**, shedding light on collaborative dynamics. A notable observation is the prominent role played by the Universidad Del Rosario, which engages in the highest number of collaborations among all institutions. This outcome

naturally stems from its leadership position within the Alianza EFI.

Moreover, by examining the node degree of each institution in the graph, it becomes apparent that the majority of institutions collaborate with up to three other entities. There are only a few instances where more than three connections exist, as depicted in Figure **3-6**. By analyzing these findings, it can be deduced that there is an opportunity to enhance collaborations among Alianza EFI member institutions beyond the Universidad del Rosario.

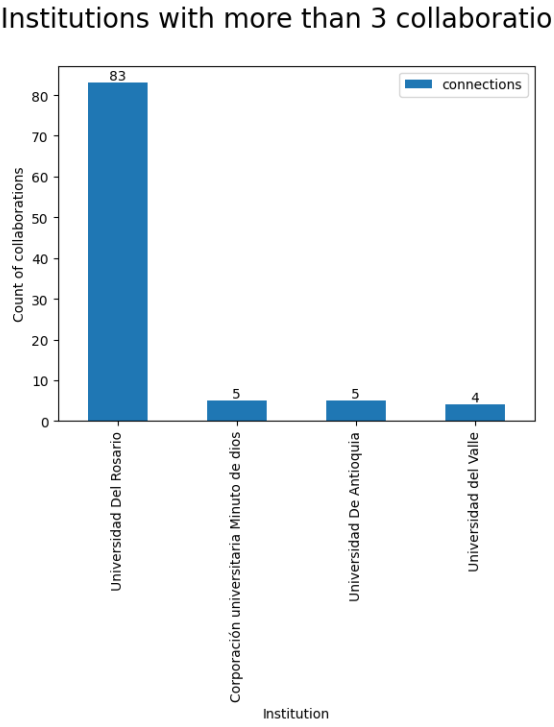


Figure 3-6: Institutions with more than three collaborations.

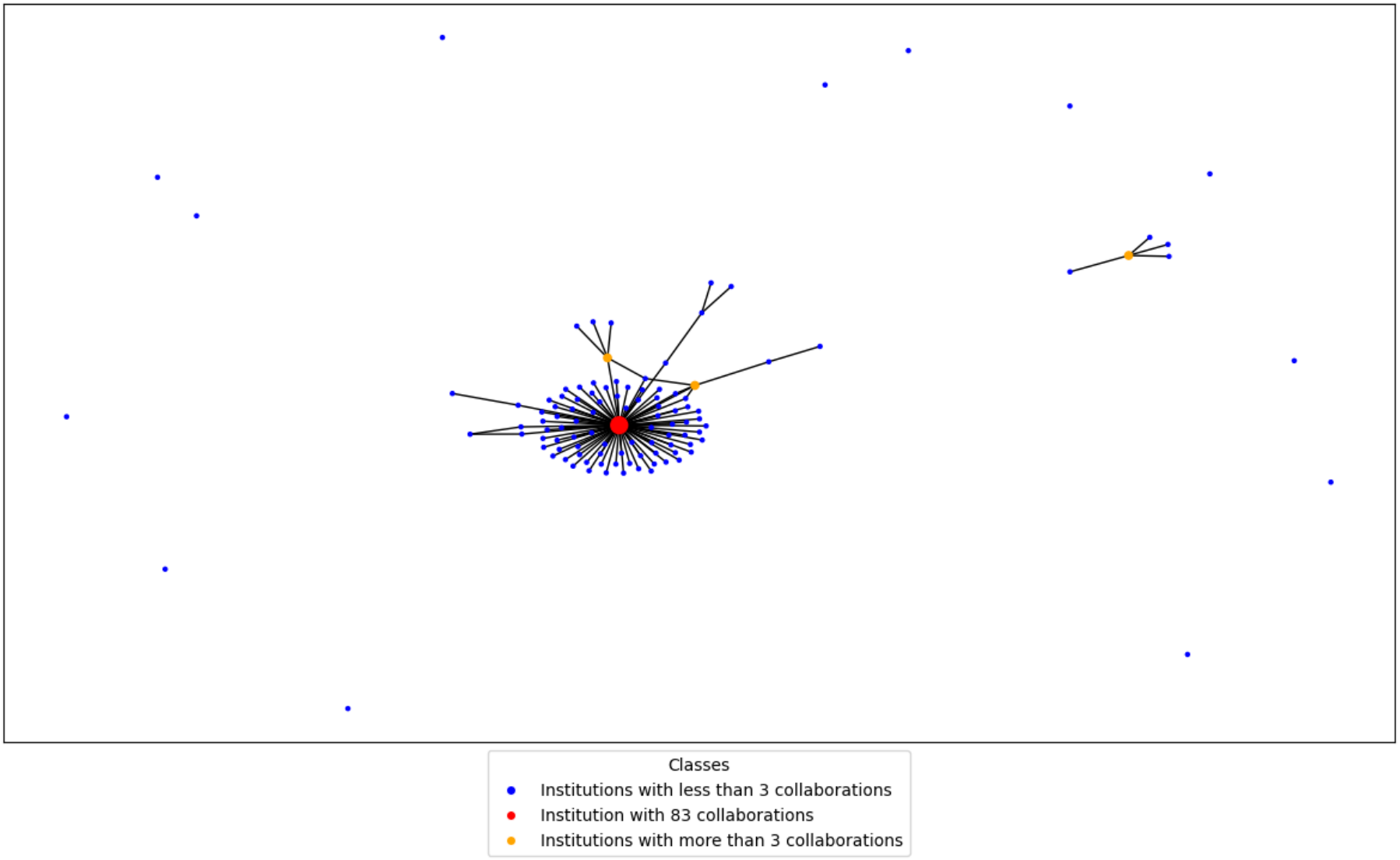


Figure 3-7: Graph representation of the co-authorship network by institution of the Alianza EFI project authors. Red: Universidad del Rosario with 83 collaborations; Orange: institutions with more than 3 collaborations; blue: institutions with less than 3 collaborations

.

3.3 Graph machine learning in action

The subsequent section presents the achieved outcomes from the application of graph learning algorithms for tasks such as node prediction, link prediction, and community detection. The Python implementation involves the creation of an undirected graph structure using NetworkX. This graph is then exported to PyTorch Geometric to facilitate the utilization of various machine learning embedding techniques. After the graph export, the resulting data object for PyTorch Geometric possesses the following attributes:

```
Data(x=[390, 101], edge_index=[2, 636], y=[390],
num_classes=7, train_mask=[390], val_mask=[390],
test_mask=[390]),
```

where `x`:= a tensor with dimensions (390, 101), representing 390 nodes, each with 101 attributes; `edge_index`:= the link information, with the number 2 indicating the columns of the edge list representation (i.e. (origin, destination)). The number 636 represents the quantity of links, accounting for the fact that the graph is undirected and counts each link twice; `y`:= labels of each node; `num_classes`:= the number of classes, which in this case is 7 corresponding to the Alianza EFI projects; `train_mask`:= corresponding to the nodes labeled as the training set; `val_mask`:= corresponding to the nodes labeled as the validation set; `test_mask`:= corresponding to the nodes labeled as the test set.

It is important to note that the division into different sets does not disrupt the graph's structure but rather assigns different labels to nodes for the purpose of training, validation, and testing.

3.3.1 Grahp statistics

The following are statistics describing the co-authorship network:

- Number of nodes: 390;
- Number of edges: 318;
- Average degree: 1.6, which indicates a low conectivity;
- Clustering coefficient; 0.02, which indicates a low tendency to form communities.

The top five authors with the highest centrality degree, indicating the most connected individuals, are:

- Darwin Cortes, 0.084,
- Paul Rodriguez, 0.069,

- Cesar Mantilla, 0.056,
- Juan Miguel Gallego, 0.041,
- Mariana Blanco, 0.041

The top five authors with the highest betweenness centrality, which signifies the most influential nodes that act as bridges between other nodes, are:

- Paul Rodriguez, 0.088,
- Cesar Mantilla, 0.067,
- Darwin Cortes, 0.062,
- Juan Miguel Gallego, 0.040,
- Maria Elvira Guerra-Cujar, 0.037

The top five authors with the highest closeness centrality, indicating nodes that are easily accessible from any other node, are:

- Paul Rodriguez, 0.144,
- Santiago Ortiz, 0.135,
- Cesar Mantilla, 0.130,
- Diana Londono, 0.128,
- Alvaro Morales, 0.127

3.3.2 Node classification

The implementation for node prediction can be summarise in four steps: Model architecture, training function, evaluation function and model training and evaluation.

Model architecture For node prediction, four different architectures were compared: MLP, GCN, GraphSAGE, and GAT. Generally, these architectures consist of multiple layers, including graph convolutional layers, aggregation layers, activation functions, and output layers. As shown in Figure **3-8**, the architectures are limited to three hidden layers to avoid over-smoothing [14]. In all the implemented architectures, the input dimensions of the first layer correspond to the number of node features. The output layer produces the final predictions, which correspond to one of the seven Alianza EFI projects. To achieve this, a Log-Softmax activation function is applied to obtain the class classification.

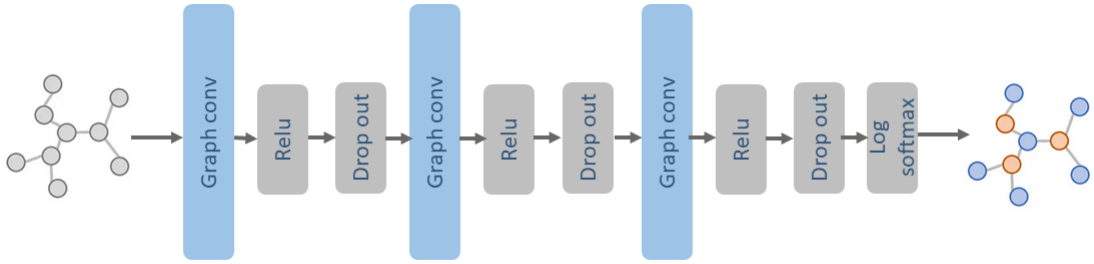


Figure 3-8: Schematic representation for the architecture network implemented for node prediction. Figure shows the case for GCN, but the structure remains the same for GrahpSAGE and GAT. Graph icons taken from [15].

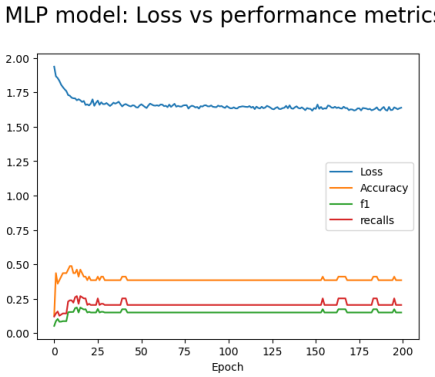
The results obtained indicate that, for the specific working graph, the GCN model achieved the highest performance, with an accuracy of 55.1%. A summary of the performance of all the tested models is provided in Table **3-5**. Additionally, the visualization depicting the evolution of loss across epochs is presented in Figure **3-9**. Notably, the MLP exhibited the lowest performance, likely attributed to its consideration of only node features and not edge information. This comparison underscores the substantial enhancement that edge information brings to node prediction tasks within graphs.

It is important to highlight that each node's parameters include the statistical values described of centrality, node degree and clustering coefficient, described in Chapter 2. These values enrich the graph's information and contribute to the training of the models. The fact that the MLP model performs the worst among the models can be attributed to its lack of information regarding graph edges. This limitation prevents the MLP model from capturing the interconnectedness and collaborative patterns present in the co-authorship network, which significantly impact its predictive performance.

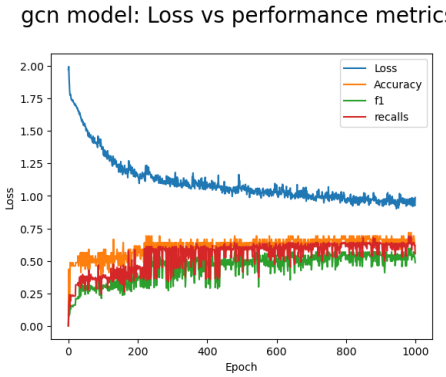
Model	MLP	GCN	GraphSAGE	GAT
Accuracy	28.2%	55.1%	47.4%	47.4%
F1-Score Macro	16.4%	42.1%	35.3%	37.2%
Recall	23.8%	44.5%	37.8%	40.6%

Table 3-4: Performance metrics comparison for different architectures to perform node prediction.

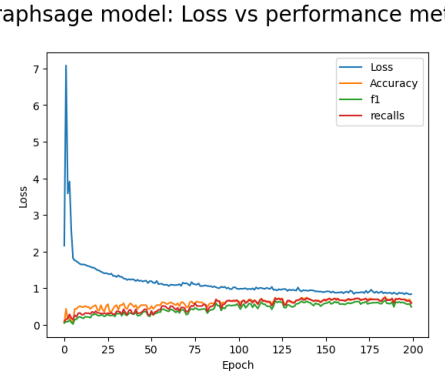
The selection of the parameters for each model was meticulously carried out using a grid search approach. This technique involved exhaustively exploring various combinations of hyperparameters to determine the optimal configuration that would lead to the best model performance. By systematically tuning parameters such as learning rate, hidden layer sizes, and dropout rates, the grid search ensured that the models were fine-tuned to their highest



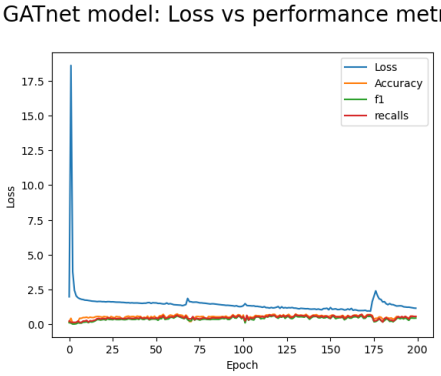
(a) Loss vs accuracy for MLP embedding



(b) Loss vs accuracy for GCN embedding



(c) Loss vs accuracy for GraphSAGE embedding



(d) Loss vs accuracy for GAT embedding

Figure 3-9: Count of authors and products per Alianza EFI project.

potential. This rigorous process allowed us to identify the most effective settings for each architecture, ensuring that the comparison among models was fair and reflective of their true capabilities.

Finally, one of the valuable applications of generating the node prediction model, specifically the GCN model, lies in its potential use for new authors who join the Alianza EFI community. By analyzing the texts and papers produced by these new authors, the model can assist in identifying the Alianza EFI project that aligns most closely with their research interests. This application offers several advantages. Firstly, it streamlines the process of onboarding new members, helping them quickly integrate into the research community. Secondly, it provides personalized recommendations that enhance collaboration and engagement, ensuring that each author contributes meaningfully to projects aligned with their expertise. Lastly, this approach can contribute to the overall efficiency of the research network by strategically matching authors to projects, thereby maximizing the impact of their contributions and fostering a more cohesive research ecosystem.

3.3.3 Link prediction

The implementation for link prediction drew inspiration from [15], and its schematic representation can be observed in Figure **3-10**. This approach can be comprehensively explained through the following steps:

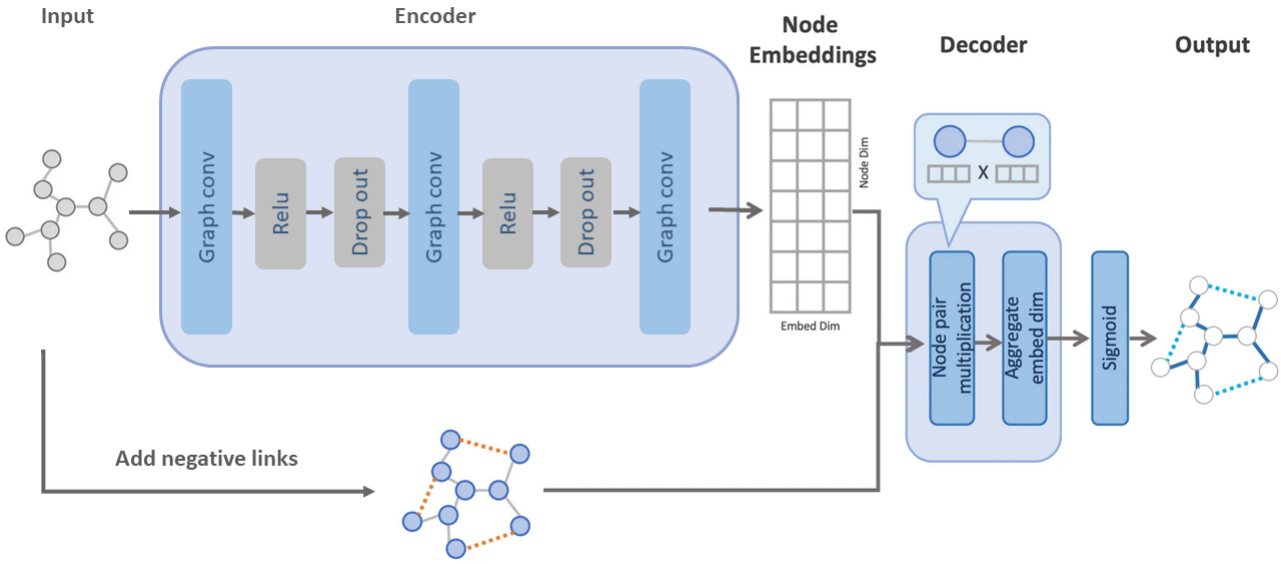


Figure 3-10: Schematic representation for link prediction network. Modified from [15].

Encoding Phase Given that the GCN model exhibited the best performance, it was chosen to create node embeddings. The GCN architecture comprises multiple GCN convolutional layers, limited to a maximum of three hidden layers. Activation functions, such as ReLU, were applied following each convolutional layer to introduce non-linearity. Dropout regularization was incorporated to mitigate overfitting during training.

Decoding Phase The decoding phase focused on predicting the presence or absence of edges between nodes. The node embeddings obtained from the encoding phase were used to compute edge probabilities. The edge scores were computed by applying a dot product to pairs of node embeddings corresponding to the candidate edges. The computed edge scores were then transformed using a sigmoid function to obtain the edge probabilities [14].

Training and Evaluation Model training utilized a dataset that comprised both positive edges (existing edges) and negative edges (non-existing edges). Negative sampling was employed to balance the positive and negative samples during training. The training loss was

computed using cross-entropy loss, comparing the predicted edge probabilities with actual ground truth labels. Model parameters underwent optimization through an Adam optimizer, set with a learning rate of 0.01 and a weight decay of $5e-4$. Training spanned 200 epochs, with loss values recorded to monitor the training's progress.

Metric	AUC	Accuracy	F1-Score Macro
Link prediction	97.4%	65.1%	74.1%

Table 3-5: Performance metrics for link prediction model.

and generates up to 10.000 new positive edges. Evolution of performance metrics are shown in Figure **3-11**.

After conducting link prediction using the implemented model, as shown in Table **3-5**, the provided performance metrics for the link prediction task offer valuable insights into the algorithm's effectiveness in predicting new edges within the original graph. The AUC (Area Under the ROC Curve) score of 97.4% indicates that the model has a high discriminatory power in distinguishing between positive and negative edge predictions. This suggests that the model is proficient at ranking positive edges higher than negative ones, which is a positive attribute for link prediction tasks.

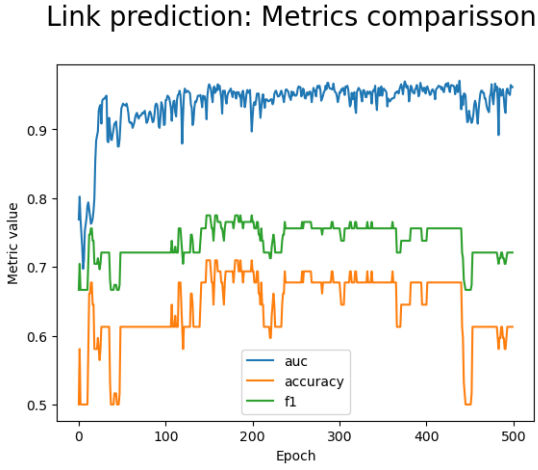


Figure 3-11: Evolution of performance metrics for link prediction model.

However, the accuracy score of 65.1% reveals that the model's overall accuracy in correctly predicting edges is moderate. While accuracy is a widely used metric, it can be affected by class imbalance and may not provide a comprehensive view of the model's performance. The F1-Score Macro of 74.1% indicates a reasonable balance between precision and recall, accounting for class imbalances. This suggests that the model achieves a satisfactory trade-off between minimizing false positives and false negatives.

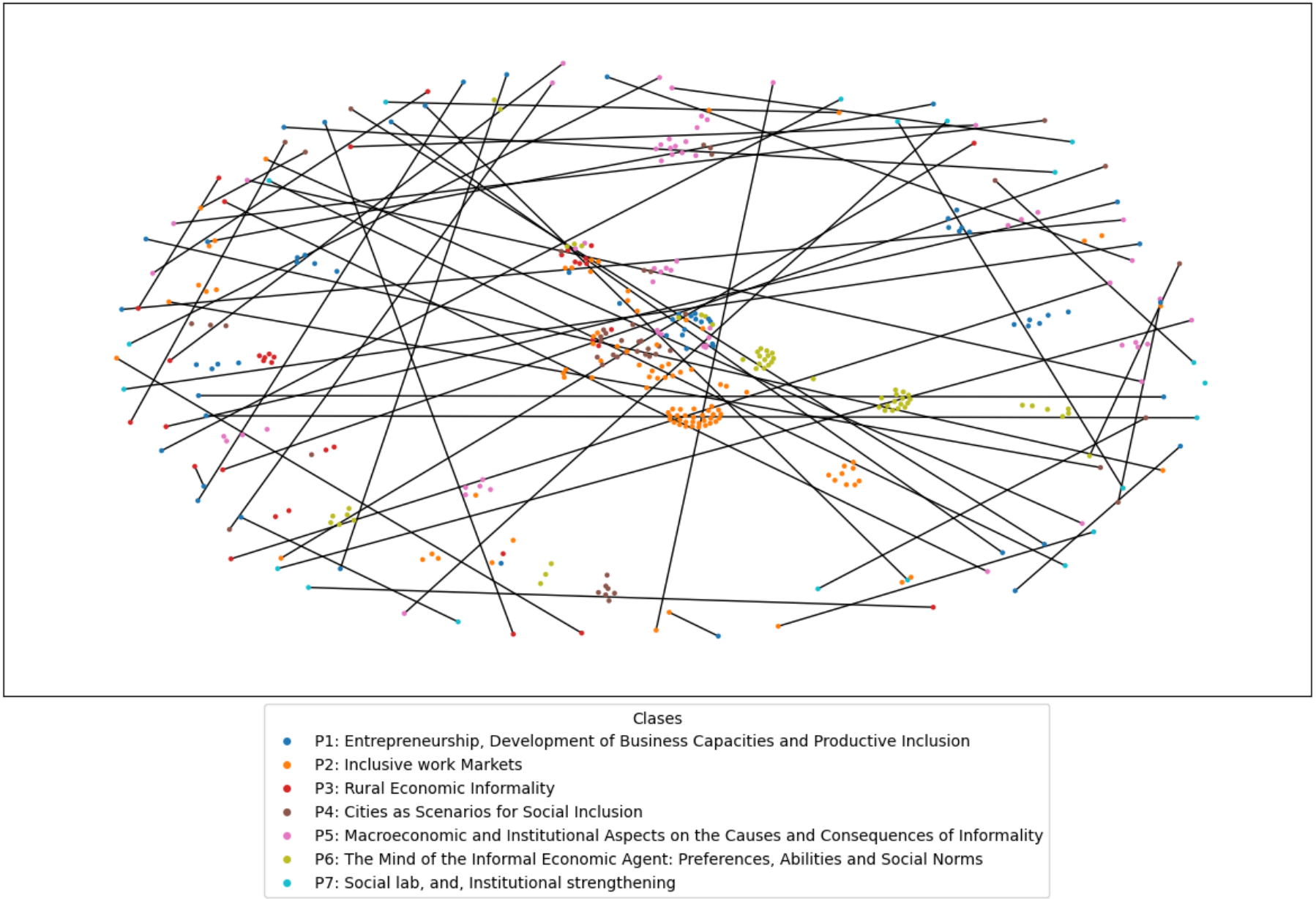


Figure 3-12: Link prediction results over the co-authorship network of the Alianza EFI project authors.

Finally, the application of the link prediction model to the authors within Alianza EFI offers an opportunity to foster new collaborations among authors who share similar research interests. By identifying potential connections between previously isolated authors, the model can facilitate the formation of partnerships that might have otherwise gone unnoticed. This not only encourages knowledge exchange but also promotes the exploration of novel research directions. The advantage of utilizing link prediction within the Alianza EFI lies in its ability to leverage the underlying network structure and predict potential collaboration opportunities, enhancing the project's overall research productivity and impact.

3.3.4 Community detection

To implement community detection in the co-authorship graph, a combined approach of GCN embedding and t-SNE (t-Distributed Stochastic Neighbour Embedding) dimensionality reduction was employed. The choice of using GCN was influenced by its superior performance in node prediction tasks. The implementation can be summarized as follows:

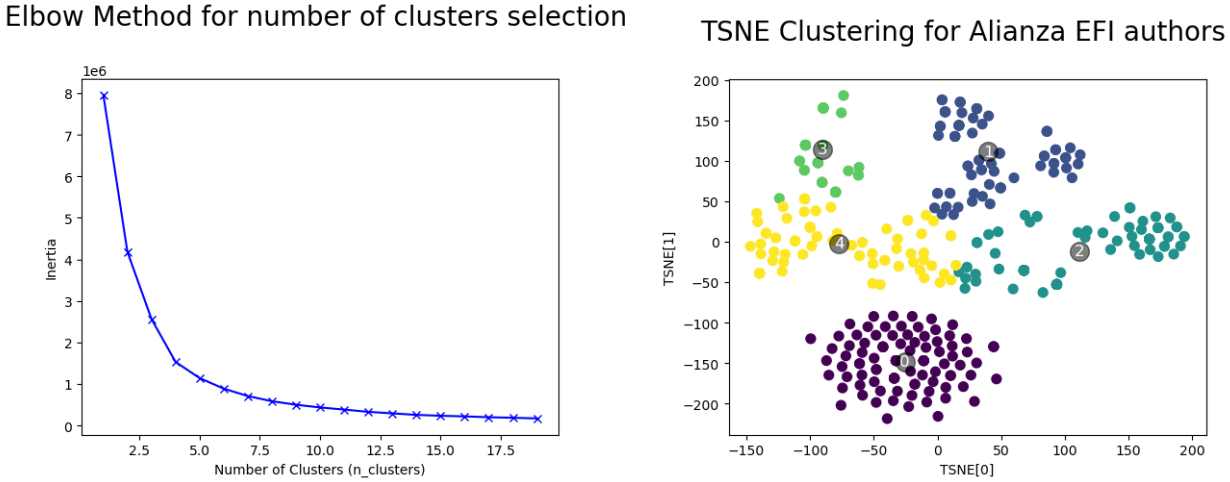
GCN Embedding A GCN model was developed to learn node embeddings that capture the graph's structural information. The approach is similar to the node prediction setup, with a distinction being that an activation function is not employed at the end of the network for classification purposes. In this case, the GCN is exclusively used to embed node information. The model consisted of up to three convolutional layers, activation functions, and pooling. The GCN output dimension was set to seven features.

t-SNE Dimensionality Reduction To visualize and cluster the acquired node embeddings, t-SNE dimensionality reduction was employed. This technique projected the embeddings into a two-dimensional space, allowing for visualization and cluster analysis.

Quantity of Clusters Determination The optimal number of clusters for the graph was determined using the elbow technique. This involved performing K-means clustering on the reduced embeddings for various cluster quantities. The optimal number of clusters was selected at the point where the inertia displayed a significant change in behavior, resembling an "elbow."

Community Detection and Visualisation Once the optimal number of clusters was identified, K-means clustering was applied to the reduced embeddings using the chosen cluster quantity, which in this case was five. Subsequently, each node was assigned to a specific cluster based on its proximity to the cluster centroid. The outcomes of the clustering were visualized in a scatter plot, with nodes represented as points and colored according to their respective clusters.

The results of this implementation are depicted in Figure **3-13**, illustrating the identification of five distinct communities numbered from zero to four. This visualization provides an insightful depiction of the different communities' formation within the co-authorship network. Additionally, the subsequent section offers a comprehensive description of the themes and participants associated with each community.



(a) Elbow method for the selection of the best number of clusters. The elbow is at $n = 5$.

(b) Communities found for Alianza EFI authors.

Figure 3-13: Community detection for the Aliza EFI authors.

Community 0 This community has 101 participants, and the topics of interest for them are:

- P1: Entrepreneurship, Development of Business Capacities and Productive Inclusion; with 27 participants,
- P7: Social lab, and, Institutional strengthening; with 20 participants,
- P5: Macroeconomic and Institutional Aspects on the Causes and Consequences of Informality; with 17 participants,
- P3: Rural Economic Informality; with 15 participants,
- P4: Cities as Scenarios for Social Inclusion; with 11 participants,
- P2: Inclusive work Markets; with 10 participants
- P6: The Mind of the Informal Economic Agent: Preferences, Abilities and Social Norms; with 1 participant.

Community 1 This community has 79 participants, and the topics of interest for them are:

- P1: Entrepreneurship, Development of Business Capacities and Productive Inclusion; with 18 participants,
- P6: The Mind of the Informal Economic Agent: Preferences, Abilities and Social Norms; with 18 participants,
- P5: Macroeconomic and Institutional Aspects on the Causes and Consequences of Informality; with 17 participants,
- P2: Inclusive work Markets; with 17 participants,
- P4: Cities as Scenarios for Social Inclusion; with 5 participant,
- P3: Rural Economic Informality; with 4 participant.

Community 2 This community has 83 participants, and the topics of interest for them are:

- P2: Inclusive work Markets; with 44 participants,
- P1: Entrepreneurship, Development of Business Capacities and Productive Inclusion; with 18 participants,
- P6: The Mind of the Informal Economic Agent: Preferences, Abilities and Social Norms; with 18 participant.
- P5: Macroeconomic and Institutional Aspects on the Causes and Consequences of Informality; with 2 participants,
- P4: Cities as Scenarios for Social Inclusion; with 1 participants,

Community 3 This community has 51 participants, and the topics of interest for them are:

- P6: The Mind of the Informal Economic Agent: Preferences, Abilities and Social Norms; with 19 participants,
- P2: Inclusive work Markets; with 15 participants,
- P3: Rural Economic Informality; with 9 participants,
- P4: Cities as Scenarios for Social Inclusion; with 5 participants,
- P5: Macroeconomic and Institutional Aspects on the Causes and Consequences of Informality; with 2 participants,
- P1: Entrepreneurship, Development of Business Capacities and Productive Inclusion; with 1 participants,

Community 4 This community has 76 participants, and the topics of interest for them are:

- P4: Cities as Scenarios for Social Inclusion; with 28 participants,
- P5: Macroeconomic and Institutional Aspects on the Causes and Consequences of Informality; with 25 participants,
- P2: Inclusive work Markets; with 15 participants,
- P3: Rural Economic Informality; with 7 participants,
- P1: Entrepreneurship, Development of Business Capacities and Productive Inclusion; with 1 participants.

Validating the detected communities within the co-authorship network reveals distinct insights into their composition and potential interactions. Community 0 primarily comprises authors identified as isolated researchers. Meanwhile, both communities 2, 3 and 4 shows well-defined structures. Particularly, community 1 is noticeable for its composition of isolated smaller sub-communities out of the inner sub-graph.

This analysis underscores several key points. First, community 0 exhibits potential for exploitation, as it highlights isolated authors who share common research interests. These shared interests could potentially catalyze collaboration and productivity within the alliance. Secondly, community 1's scattered nature of small groups presents an opportunity for integration, suggesting the potential to unify these authors to foster the creation of novel products. Finally, the mixed nature of community 4, due to its intertwining with communities 2 and 3, positions it as a bridge of interaction between diverse research areas. This could give rise to a platform for the generation of heterogeneous research topics and interdisciplinary collaboration.

In essence, the application of community detection techniques emerges as a powerful tool with the potential to cultivate new collaborations among isolated authors. By identifying shared research interests and facilitating connections between these researchers, the Alianza EFI can harness a heightened level of productivity and innovation. This proactive approach to community-driven collaboration holds the promise of amplifying the impact of the alliance's efforts, fostering a thriving ecosystem of knowledge exchange and cross-pollination, ultimately enriching the collective pursuit of advancement and discovery.

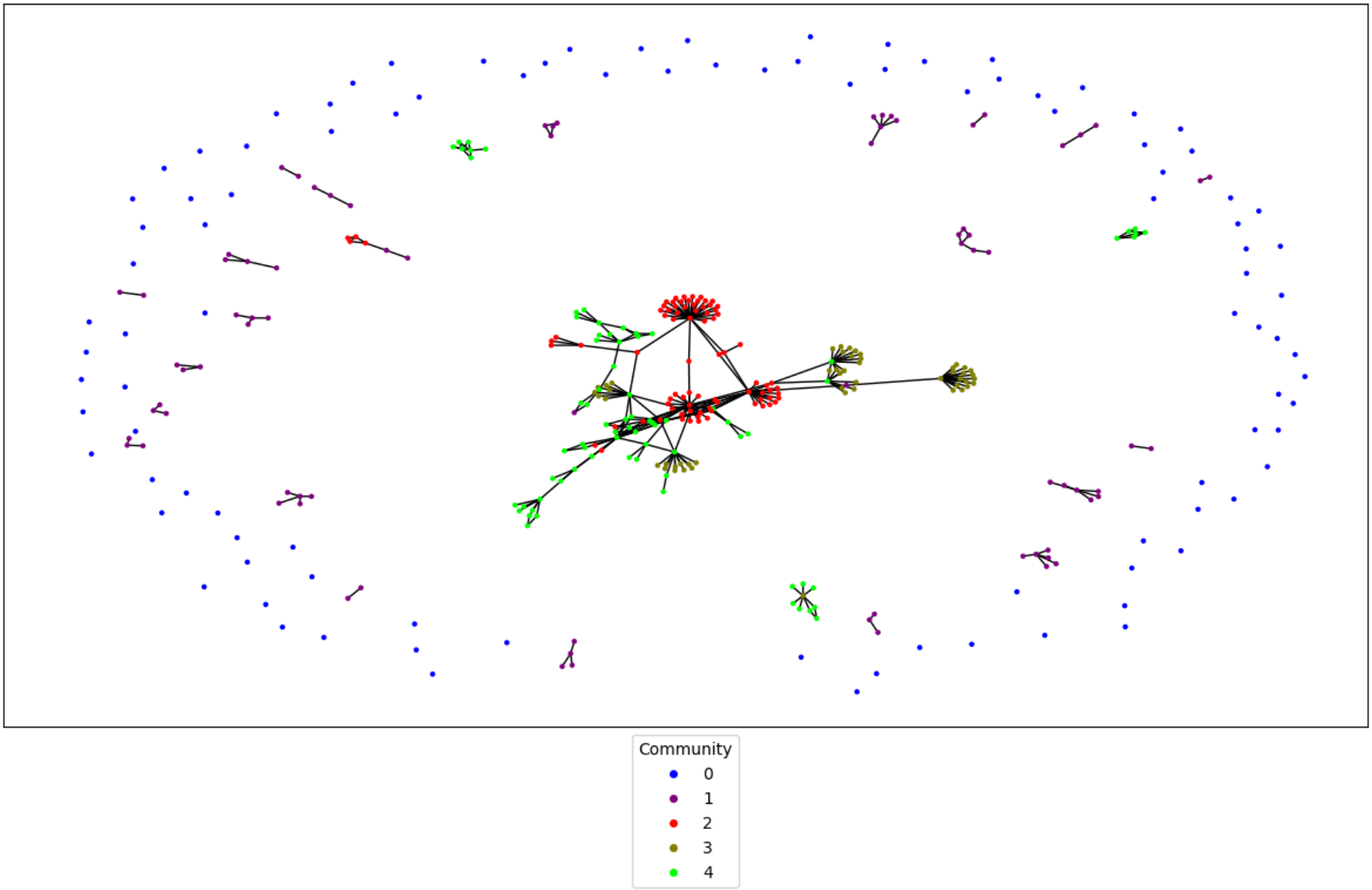


Figure 3-14: Community detection results over the co-authorship network of the Alianza EFI's project authors.

Conclusions

The conducted descriptive analysis sheds light on significant insights regarding the Alianza EFI project. It discloses that the alliance comprises 390 unique authors affiliated with 112 institutions, generating a total of 274 distinct products. Notably, the Universidad del Rosario assumes a pivotal role in institutional collaborations, engaging in 83 unique partnerships. In contrast, other alliance institutions participate in fewer than five collaborations, indicating potential opportunities for fostering greater collaboration among alliance members. Additionally, authors affiliated with the Universidad del Rosario stand out as the most prolific contributors in terms of both product generation and collaborations within the alliance.

As hypothesized, the incorporation of node statistics proves invaluable in identifying influential authors who occupy crucial roles within the alliance. Furthermore, the application of node prediction techniques facilitates the anticipation of authors' new areas of interest, thereby stimulating the creation of novel products. On the other hand, the link prediction approach emerges as a potent strategy for uncovering potential connections between authors. Notably, it effectively suggests potential links for isolated authors, fostering new collaborations and knowledge-sharing within the alliance. Lastly, community detection analysis exposes the presence of five distinct communities within the alliance. Each community exhibits specific interests and aligns with particular Alianza EFI projects. Understanding these community structures offers valuable insights into the multifaceted dimensions of the informal sector addressed by the alliance, fostering targeted collaboration and knowledge exchange among community members.

Looking ahead, it would be advantageous to extend the analysis to include external authors who reference the alliance's papers. This supplementary dimension would yield insights into the global reach and impact of the alliance's products within the international research landscape. By examining citations from authors worldwide, it becomes possible to gauge the global significance and influence of the alliance's research outcomes. This broader perspective not only enhances our comprehension of the alliance's contributions but also sheds light on the wider scientific context. Such an examination would help evaluate the alliance's visibility, influence, and potential collaborations beyond its immediate network of affiliated institutions and authors. Integrating this aspect into future research would provide a comprehensive view of the alliance's impact and facilitate the exploration of potential global partnerships and knowledge exchange on a broader scale.

References

[1] *AI in Colombia*. Available at https://oecd.ai/en/dashboards/countries/Colombia. 2022. – Accessed: 2022-11-28

[2] *Graph Neural Networks: Libraries, Tools, and Learning Resources*. Available at https://neptune.ai/blog/graph-neural-networks-libraries-tools-learning-resources. 2022. – Accessed: 2022-11-28

[3] *PyG is the ultimate library for Graph Neural Networks*. Available at https://www.pyg.org/. 2022. – Accessed: 2022-11-29

[4] *Sobre la Alianza EFI*. Available at https://alianzaefi.com/miembros/. 2022. – Accessed: 2022-11-29

[5] Antonia Creswell, Vincent Dumoulin Kai Arulkumaran Biswa S. ; Bharath, Anil A.: Generative Adversarial Networks: An Overview. En: *IEEE-SPM* (2017)

[6] Aric A. Hagberg, Daniel A. S. ; Swart, Pieter J.: Exploring network structure, dynamics, and function using NetworkX. En: *KDD '16: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (2008), p. 11–15

[7] Bacciu, Davide ; Errica, Federico ; Micheli, Alessio ; Podda, Marco: A gentle introduction to deep learning for graphs. En: *Neural Networks* 129 (2020), p. 203–221. – ISSN 18792782

[8] Barabási, Albert-László: *Network science*. Available at http://networksciencebook.com/. 2022. – Accessed: 2022-11-29

[9] Chaudhary, Mukesh: *TF-IDF Vectorizer scikit-learn*. Medium. April 2020. – Accessed: 10/06/2023

[10] Elinas, Pantelis: *Knowing Your Neighbours: Machine Learning on Graphs*. Medium. June 2019. – Accessed: 01/07/2022

[11] Garfield, Eugene: Citation analysis as a tool in journal evaluation. En: *Science* 178 (1972), Nr. 4060, p. 471–479

[12] Hamilton, William L.: Graph Representation Learning. En: *Synthesis Lectures on Artificial Intelligence and Machine Learning* 14 (2020), Nr. 3, p. 1–159

[13] Iosifidis, Anastasios: *Deep Learning for Robot Perception and Cognition*. Academic Press, 2022

[14] Leskovec, Jure: *CS224W: Machine Learning with Graphs*. Standford Online Course. 2020. – Accessed: 05/01/2023

[15] Masui, Tomonori: *Graph Neural Networks with PyG on Node Classification, Link Prediction, and Anomaly Detection*. Medium. October 2022. – Accessed: 01/08/2022

[16] Needham, Mark: *Graph Algorithms: Practical Examples in Apache Spark and Neo4j*. London : O'REilly, 2019

[17] Nestor Maslej, Erik Brynjolfsson John Etchemendy Katrina Ligett Terah Lyons James Manyika Helen Ngo Juan Carlos Niebles Vanessa Parli Yoav Shoham Russell Wald Jack C. ; Perrault, Raymond: The AI Index 2023 Annual Report / Stanford University. 2023. – Informe de Investigación. – 386 p.

[18] Newman, Mark: *Networks: An Introduction*. New York, NY : Oxford University Press, 2010. – ISBN 978–0–19–920665–0

[19] Oliveira, Edgar Thiago D. *Deep Learning and its Applications today*. May 2023

[20] Russell, Stuart J. ; Norvig, Peter: *Artificial Intelligence: A Modern Approach*. Pearson, 2010

[21] Sanchez-Lengeling, Benjamin; e.: *A Gentle Introduction to Graph Neural Networks*. Available at https://distill.pub/2021/gnn-intro/graph-to-tensor. 2021. – Accessed: 2022-11-29

[22] Sandra Cortesi, Marcelo Cabrol Alejandro Correa Urs Gasser Carol Hullin Alejandro Jaimes Malavika Jayaram Clara Mosquera López Riel M. *Recomendaciones de la misión de expertos IA*. 2022

[23] Wu, Zonghan ; Pan, Shirui ; Long, Guodong ; Jiang, Jing ; Chang, Xiaojun ; Zhang, Chengqi: A comprehensive survey on graph neural networks. En: *IEEE Transactions on Neural Networks and Learning Systems* 32 (2021), Nr. 1, p. 4–24

[24] Xia, Feng ; Sun, Ke ; Yu, Shuo ; Aziz, Abdul ; Wan, Liangtian ; Pan, Shirui ; Liu, Huan: Graph Learning: A Survey. En: *IEEE Transactions on Artificial Intelligence* 2 (2021), p. 109–127