Computer Science Project

Master 297 Digital Economics - Yuren Jin, Yujing Zhang

January 2024

1 Introduction

In the evolving landscape of academic research, understanding the dynamics of scholarly work and the connections between researchers offers invaluable insights into the development of scientific domains and the collaborative networks that underpin them. This project aims to delve into the rich repository of articles published by researchers from the Laboratoire d'Economie de Dauphine (LEDa) at Université Paris Dauphine , accessible through the HAL open archive (https://hal.science/). By leveraging data science techniques, including web scraping, topic modeling and social network analysis, our objective is to uncover the thematic evolution of LEDa's research and to map the intricate web of author collaborations over time.

1.1 Description of the data sets

To acquire a comprehensive dataset of publications from the Laboratoire d'Economie de Dauphine (LEDa), we leveraged the capabilities of the HAL open archive's API. This digital platform offers access to a vast repository of scholarly articles across various disciplines. This comprehensive repository provides open access to a wide range of academic outputs, including journal articles, conference papers, and working papers authored by LEDa researchers.

1.1.1 API Query Configuration

The endpoint used for our query was https://api.archives-ouvertes.fr/search/, tailored to fetch records relevant to our research objectives.

To ensure the precision and relevance of our data retrieval, we structured our query with specific parameters:

- Search Term (q): 'LEDa Dauphine' as our search keywords.
- Fields to Retrieve (fl): Our query requested specific fields for each publication, including the document ID (docid), the publication label (label_s), and the document's URI (uri_s). These fields were selected to provide essential identifiers and access links for each record, facilitating further analysis and referencing.
- Starting Record (start): The starting point of our data retrieval to the first record (0).

- Number of Records (rows): To capture an extensive dataset, we specified a retrieval of 10,000 records. This was adjusted to ensure a broad coverage of LEDa's publication history, but the truth is we only get 1746 rows.

1.1.2 Extracting Metadata

Our next critical step involved extracting and structuring key metadata from the raw data to facilitate further analysis. This phase aimed to parse out the authors, paper titles, and publication years from the combined label field provided in the dataset.

To accomplish this, we designed a Python function extract_info, leveraging regular expressions (regex) to accurately identify and separate the author names, paper titles, and publication years embedded within the label field of each record. The regex pattern used was:

```
^(.*?)\. {1,2}(.*?)\.\s*(.*?)(\d{4})
```

After the extraction and structuring process, we added the newly created columns (author, title, year) to verify the accuracy of the information extracted and to provide a clear overview of the dataset's enriched composition.

1.1.3 Web Scraping for Abstract

In topic modeling, the richness and depth of the textual data under analysis significantly impact the quality and interpretability of the resulting topics. To enrich our dataset with detailed content for a more nuanced topic modeling process, we developed a Python function, scrape_abstract, designed to extract abstracts from academic publications hosted on web pages.

The function operates on the following principles:

- HTTP Requests: Utilizing the requests library, the function initiates a GET request to the specified URL

- Status Code Verification: Upon receiving a response, the function checks the status code to ensure a successful connection (HTTP 200 OK).

- HTML Parsing: Leveraging the BeautifulSoup library, the function parses the HTML content of the web page.

- Abstract Extraction: The core of the function searches for the div element tagged with the class abstract-content active and the language attribute set to either English (en) or French (fr).

- Exception Handling: The function is fortified with exception handling to manage potential issues arising from network problems or other errors during the request process. By catching RequestException, the function returning None to indicate the absence of retrievable abstract text.

Our final dataset comprises 1,746 entries, organized across 7 columns, which include crucial metadata such as author, title, year, and abstract. A notable aspect of our dataset is the presence of 777 'None' values for the abstract field, indicating instances where abstract texts could not be retrieved. However, the collection of 969 abstracts represents a valuable corpus for deep textual analysis.

2 Project Overview

Our objective is to conduct descriptive analysis to provide a comprehensive overview of LEDa's research publications, mapping out the landscape of research themes, authorship patterns, and publication trends over time.

2.1 Network Analysis

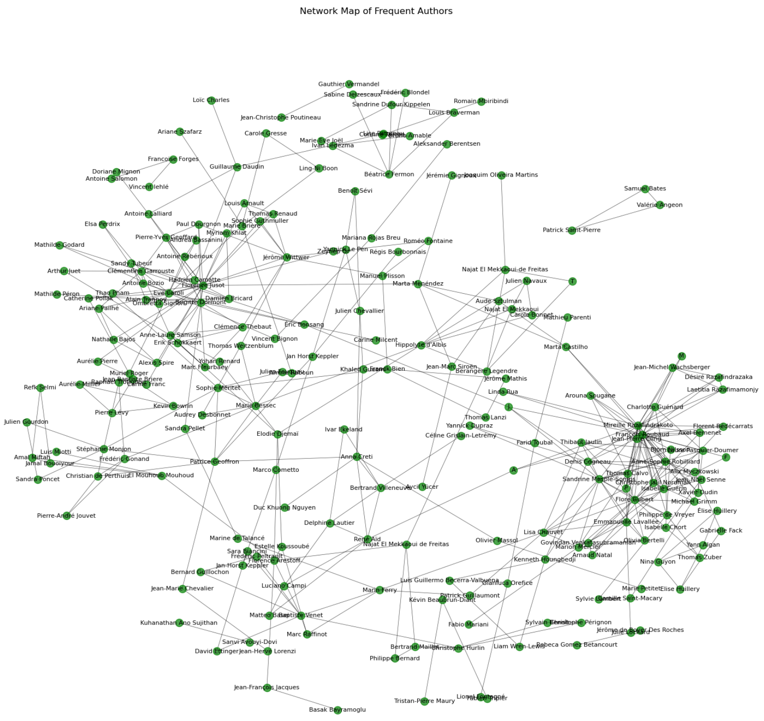

Figure 1: Network Map of Frequent Authors

Network analysis stands as a pivotal component of our project, aiming to dissect the collaborative landscape within the LEDa. This analytical framework will enable us to visualize and quantify the complex web of academic collaborations. Through this aspect, we seek to uncover the patterns of co-authorship, identify central figures within the network, and explore the formation of research clusters.

The initial step involved conducting a frequency analysis of author appearances within the dataset. This process utilized Python's Counter from the collections module to tally the number of publications attributed to each author. The rationale behind this was to identify authors with substantial contributions, under the premise that more frequent collaborations are indicative of stronger or more central roles within the research network.

Next, leveraging the networkx library, we proceeded to construct a graph representation of the filtered co-authorship network (Figure 1). The criteria for inclusion in this network were set to authors who appeared more than a predefined threshold of times, specifically chosen as three instances of collaboration, to ensure that the network encapsulated meaningful academic interactions. For each publication in the dataset, pairs of authors were identified and filtered based on the aforementioned frequency criterion. An edge was then added between every pair of these frequent collaborators, signifying a co-authorship link.

2.1.1 Degree Distribution in the Co-Authorship Network

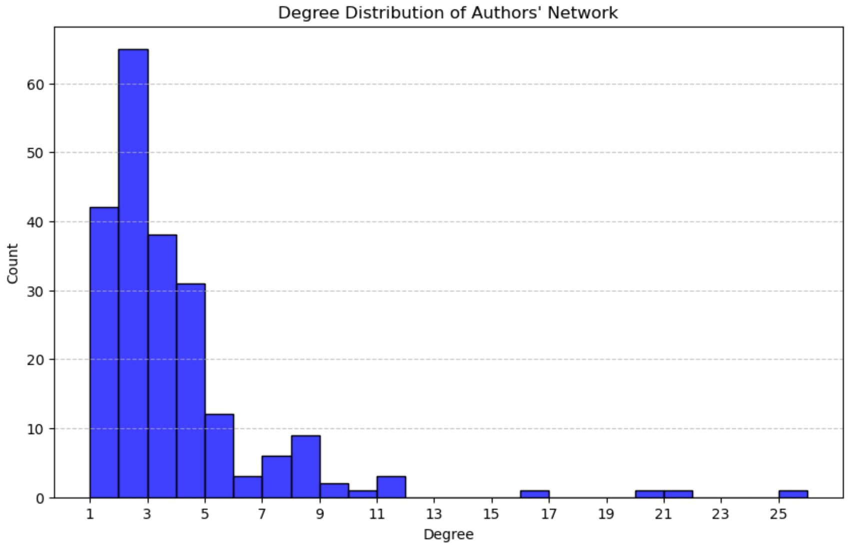

Figure 2: Degree Distribution

Following the construction of the filtered co-authorship network, a crucial step in our network analysis involves examining the degree distribution of authors within this network. The degree of a node (author) in a network signifies the number of connections (collaborations) that node has with others, providing insights into the collaborative patterns and the structure of the academic community.

The histogram (Figure 2) visualizes the degree distribution of the authors' network , reflecting how many collaborations each author has. The histogram shows a right-skewed distribution, which is typical for social networks. A large number of authors have a small number of collaborations (low degree), while a few authors have a large number of collaborations (high degree).

The most common degrees appear to be in the lower range (between 1 and 5), which suggests that most authors in the network tend to collaborate with only a few other authors. The authors with higher degrees (towards the right end of the histogram) suggests that there are central figures within the network. These individuals likely act as collaborative hubs, having numerous co-authorships with different researchers.

For the LEDa research community, the degree distribution could indicate how collaborative efforts are distributed among researchers.The overall shape of the distribution can also give insights into the network's connectivity. A network with many low-degree authors and few high-degree authors have a star-like structure with several sub-networks centered around highly collaborative individuals. A collaborative environment might be suggested by a more uniform distribution, while a highly skewed distribution could indicate that efforts to foster collaboration among more researchers may be beneficial.

2.1.2 Clustering Analysis

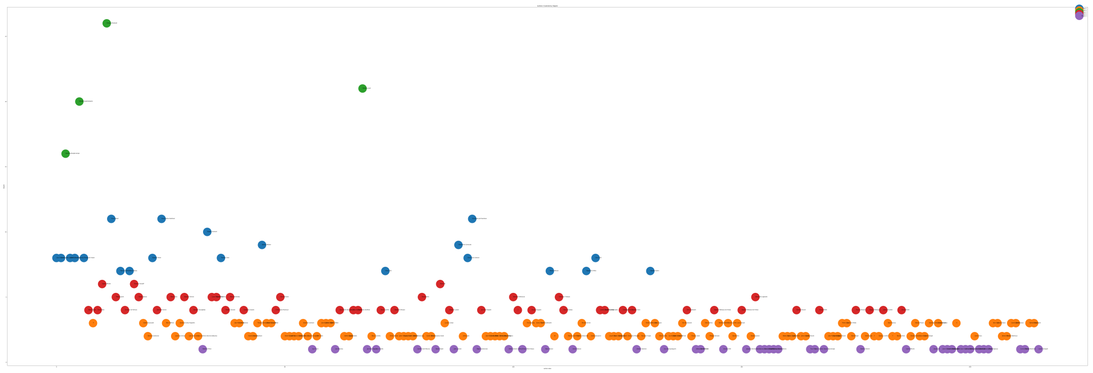

Figure 3: Authors Clustered by Degree

We began by extracting the degree of each author in the network—represented as the number of collaborations—thereby quantifying their connectivity within the scholarly community. This information was compiled into a dictionary and then transformed into a NumPy array, which facilitated subsequent analytical processing.

We leveraged the KMeans algorithm from sklearn.cluster, we applied a clustering algorithm to the array of degrees. We opted for five clusters (n_clusters=5).

While the histogram provides a univariate distribution of degrees across all authors, the scatter plot offers a bivariate distribution that separates authors into clusters. The histogram may show a concentration of authors with a lower degree, which corresponds to the larger clusters at the bottom of the scatter plot.

The scatter plot also helps identify outlier authors who have a significantly higher degree than others, which might not have been as evident in the histogram. These authors can be seen as isolated points above the main cluster groups and represent influential researchers within the network. (i.e François Roubaud, Florence Jusot, Mireille Razafindrakoto, Sandrine Meslé-Somps)

2.2 Topic Modeling

2.2.1 Language Segmentation

A critical step for topic modeling was to differentiate and preprocess texts based on their language. Given the bilingual nature of LEDa's research output, we aimed to segregate the English and French publications to tailor the topic modeling process to the linguistic nuances of each corpus.

We utilized the langdetect library, we implemented a function to detect the language of each publication by combining the title and abstract into a single text string. Publications were then classified as 'English' or 'French' based on the detected language, resulting in two distinct datasets for further analysis

2.2.2 Text Preprocessing and Modeling with LDA

A robust text preprocessing pipeline was set up, leveraging the CountVectorizer from the scikit-learn library to convert the textual data into a document-term matrix (DTM). The vectorization process was refined with custom stop words, integrating both English and French stopwords to filter out common, non-informative words from each language set. The vectorizer was also configured with parameters max_df and min_df set to exclude terms that are too common or too rare.

Then, with the preprocessed data, we applied Latent Dirichlet Allocation (LDA), a generative probabilistic model, to identify latent topics within the English and French datasets. The LDA model was configured to uncover five distinct topics from each dataset, providing a granularity that balances thematic diversity with interpretability.

2.2.3 Displaying Topics

Topic	Keywords
1	trade, informal, care, formal, results, french, economic, cost, sector, using
2	tax, aid, model, debt, show, firms, market, paper, countries, growth
3	risk, price, model, equilibrium, data, market, show, impact, france, french
4	health, social, data, household, survey, inequality, migration, countries, income, results
5	market, costs, financial, countries, energy, cost, data, new, trade, price

Table 1: English topics and their associated keywords

Topic	Keywords
1	travail, marché, entre, plus, assurance, personnes, france, économique, publics, pratiques
2	santé, soins, énergétique, non, cette, complémentaire, plus, émissions, effet, recours
3	prix, mondialisation, sociale, plus, cette, crise, pays, mondiale, accord, politique
4	plus, risque, pays, impact, entre, emploi, médecins, analyse, crise, être
5	plus, développement, pays, économique, politiques, cette, vie, comment, ans, cours

Table 2: French topics and their associated keywords

The topics encapsulates the thematic essence extracted from LEDa's research publications. The English topics could indicate a multifaceted approach to economic, health, and social issues, with a strong analytical and data-driven foundation. The

French topics highlight a more socio-economic and policy-driven orientation, with significant attention to labor, health, and globalization.

3 Difficulties encountered

During the Data collection step, primarily from web scraping and APIs, posed its own set of difficulties. Slight variations in webpage structures often led to incomplete data retrieval, necessitating a preprocessing strategy to clean and standardize the collected data. The inherent limitations of automated scraping, such as rate limits and the need for dynamic adaptation to website changes, also complicated the process. Incomplete metadata also creates difficulties in potential time series analyses, e.g. obvious errors in the year column (0352, 1007, 2030...) Increased complexity in the data pre-processing phase.

One of the primary challenges for the network analysis is the construction of the co-authorship network faced scalability issues, with the visualization of large networks particularly. Striking a balance between detail and clarity in network graphs required iterative refinement.Plus, it is complicated by the need to accurately parse author names and handle instances of multiple authors or variations in name representation.

Another challenge stemmed from the multilingual nature of the dataset. LEDa's publications are in both English and French, requiring nuanced language processing to ensure accurate analysis. Language detection and the subsequent segregation of the dataset introduced additional steps and potential points of error. Moreover, the differentiation of stopwords and linguistic subtleties between the two languages added a layer of complexity to the text processing and topic modeling phases. Even though I've implemented nltk's French stopwords and adjusted max_df and min_df, meaningless words like plus, être, cette still appear in French topic.

4 Methodology

The investigation into the research publications of the Laboratoire d'Economie de Dauphine (LEDa) was structured around a multidisciplinary methodological framework. Our approach combined techniques from data science, natural language processing (NLP), and network theory to unearth and understand the thematic and collaborative patterns within the LEDa academic community.

Data Collection Methodology:

- API Data Retrieval: We initiated the project by harnessing the HAL open archive's API to systematically collect metadata on LEDa's publications. The API allowed us to filter and retrieve data based on specific query parameters aligned with our research focus, such as the names of authors, publication titles, abstracts, and keywords.

- Web Scraping: Given the limitations of the data retrieved through the API, such as missing abstracts, we incorporated web scraping to enrich our dataset. The web scraping procedure was designed to navigate the structure of the publication web pages and extract detailed content, particularly the abstracts, which are pivotal for in-depth text analysis.

- Language Detection: We employed the langdetect library to segregate publications into English and French subsets. This preliminary step ensured that all subsequent analyses were linguistically coherent.

Analytical Methodologies:
Network Analysis:

- Co-Authorship Network Construction: A co-authorship network was constructed using the networkx library to visualize and analyze the collaborative ties within LEDa. This network mapped researchers to nodes and joint publications to edges, creating a graph representing the intricate web of academic collaborations.

- Degree Analysis: The degree of each node (author) was calculated to gauge the extent of each researcher's collaborative engagements. We plotted a degree distribution histogram and a scatter plot to understand the network's structure and to identify both central figures and peripheral participants in the research community.

- Clustering: We explored clustering algorithms to categorize authors based on their collaborative activity. K-means clustering was applied to the authors' degree data to segment the network into groups of researchers with similar levels of connectivity.

Topic Modeling:

- Text Preprocessing: We preprocessed the textual content of the publications using NLP techniques. This included combining titles and abstracts into a unified text field, tokenization, and the removal of stopwords in both English and French.

- Latent Dirichlet Allocation (LDA): To extract prevalent themes from the publications, we applied LDA, a generative statistical model that identifies topics within a corpus. We experimented with different numbers of topics and evaluated the coherence and distinctiveness of the resulting topics.

5 Conclusion

In conclusion, our project provides a data-driven perspective on LEDa's research output, highlighting the laboratory's thematic diversity and collaborative patten. The methodologies applied have proven effective in extracting meaningful patterns from academic data. We foresee that with continuous improvements and adaptations, future models will yield even greater insights, furthering our understanding of the dynamics of researches in different fields.

From the perspective of methodology, the utilize of web scraping methodologies significantly enhanced our dataset, particularly by augmenting incomplete records with abstracts directly sourced from web pages. This addition provided a richer textual corpus for the topic modeling, contributing to more nuanced and informed topic generation. The network analysis revealed a complex tapestry of collaboration, marked by a few highly connected individuals and numerous researchers with fewer collaborative ties. The degree distribution was right-skewed, typical of social networks, indicating a core-periphery structure with key figures facilitating cross-disciplinary interactions. The topic modeling yielded distinct themes that characterized LEDa's research focus where LDA model successfully captured the thematic diversity, demonstrating the breadth of LEDa's scholarly endeavors.

5.1 Future Work

Future network models could be refined by incorporating more dynamic network measures and possibly integrating temporal aspects to capture the evolution of collaborations over time. (For example, the last two blocks of the code retrieved the topic variations for all years of published articles and deposited them in a .txt file, but did not reach the point of further analysis.) For LDA, we anticipate improved performance through the incorporation of more advanced NLP techniques, such as word embeddings or topic coherence measures, to enhance the distinctiveness of the topics extracted from new data. And interpretation of the results of topic models could be subjective. Moving forward, collaborating with professors could add depth to the interpretation of topics, ensuring that the model outputs align with expert understanding of the research domains.

While our models have been tailored to LEDa's dataset, we expect that with some adjustments, they can be generalized to other academic corpora. The scalability of our methods allows for handling larger datasets, with the potential to uncover broader trends across different research fields or institutions.