IADIS International Journal on Computer Science and Information Systems
Vol. 20, No. 1, pp. 90-104
ISSN: 1646-3692

FROM DATA COLLECTION AUTOMATION TO SEQUENCE ANALYSIS: A GENERAL METHODOLOGY ENABLING DATA ANALYSIS OF CORPORATION CAREERS

Yinglei Han[1], Dario Colazzo[1] and François-Xavier Dudouet[2]
[1]CNRS, LAMSADE, Université Paris Dauphine – PSL Paris, France
[2]CNRS, IRISSO, Université Paris Dauphine – PSL Paris, France

ABSTRACT

In this paper, we outline a general methodology for analysing corporate career data using data extracted from professional CVs. Our process begins by collecting resumes in JSON format automatically and storing them in a centralized repository. These data are then systematically structured into both relational and graph databases, enabling rich, multidimensional analysis. Through the integration of database management techniques and machine learning tools, we can conduct further analysis, such as sequence analysis, to trace and interpret career paths, uncovering career patterns of business elites. This framework not only facilitates detailed empirical investigation but is also designed to be reproducible and adaptable, serving as a foundation for future research.

KEYWORDS

Corporate Career Analysis, Top Managers, Relational Database, Graph Database, Sequence Analysis

1. INTRODUCTION

Big businesses play an increasingly important role in our world, making it crucial to understand who leads them and how. More and more often, the leaders of large corporations attain their positions through their careers rather than their wealth. To better grasp the mechanisms of power within these corporations, it is essential to study their career systems, which often reveal shared career paths. However, significant challenges hinder such analysis (Koyuncu et al., 2017). First, the collection, organization, cleaning, and automatic enrichment of web-based career data remain formidable tasks for social science researchers, who frequently rely on time-consuming manual methods. Second, the lack of publicly available database architectures in this domain limits opportunities for shared learning, replicability, and scientific accumulation. This creates a gap in systematic and scalable approaches to studying corporate career data. To address these issues, we propose a comprehensive model for automating the collection, storage, and cleansing of career-related data in the social sciences. We also introduce a database architecture designed specifically for processing and analysing the career paths of corporate managers. For the sake of simplicity, this paper is not in the context of massive data. Instead, we focus on detailed descriptions of our workflow, as in the future, each step implemented in the workflow can be scaled up today by solutions (i.e., Hadoop, Spark, etc.) to cope with big data analysis.

We develop a reproducible approach, intended to serve as a replicable model for future social scientists for work including CV analysis, career analysis, and so on. Specifically, we create one workflow framework (Figure 1) for corporate manager career data analysis. To start, the workflow includes a JSON CV store to save new CV input. Following information reconstruction, we create a relational database in PostgreSQL and a graph database in Neo4j. By means of these two databases, we can conduct further analysis such as sequence analysis, network analysis, ad-hoc analysis, etc.

A previous version of this work appeared in IADIS 2025 (Han, Colazzo and Dudouet, 2025). With respect to that version, this paper includes additional content related to a more detailed description of the data collection process and of the relational database creation, as well as new content related to the construction of a graph database starting from the relational database we have designed, which paves the ways to new kinds of network analysis.

The paper is structured as follows. Section 2 reviews related work. We briefly discuss systematic career analysis in the field of top managers, including data preparation methods and sequence analysis, serving as one way of further analysis. Section 3 describes our data source (i.e., top managers from CAC 40 corporations in 2019 in our paper) and data collection method in this paper. We present how we create a relational database in Section 4, where we discuss how to conceptualise career trajectories collectively and introduce a term called “Arena” to conceptualise executive and non-executive positions. In Section 5, we describe a combination of rule-based and machine-learning methods to label career hierarchy “Rank” in our database. Section 6 introduces our graph database implemented in Neo4j. Section 7 presents the sequence analysis, demonstrating how the database can be used for further analysis. This section also includes the findings from our cluster analysis of career paths. Finally, Section 8 provides a summary of our approach and outlines directions for future work.

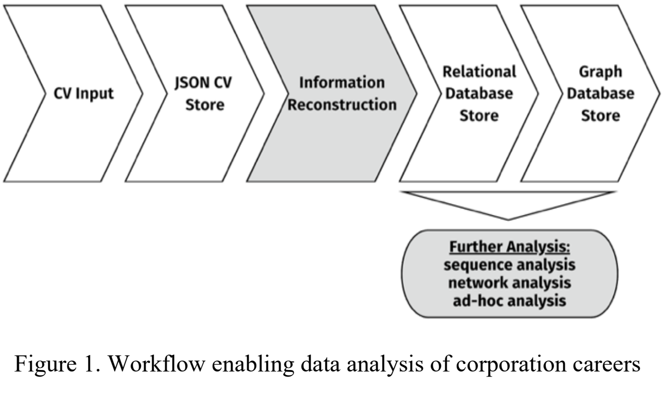

Figure 1. Workflow enabling data analysis of corporation careers

2. RELATED WORK

Elites exert disproportionate control over diverse social resources, including economic, social, cultural, political, and knowledge capital (Khan, 2012). Thus, the study of elite recruitment is crucial (Reeves et al., 2017). Recent literature has systematically explored career progression patterns and social relations of elites from a collective perspective. For example, Koch et al. (2017) analysed the career moves of Fortune 100 CEOs across various dimensions of job mobility, such as status, function, and employer. Davoine and Schmid (2022) reviewed key characteristics of top managers' careers, while Bühlmann et al. (2022) identified three career hubs that shape the global corporate elite. Henriksen and Seabrooke (2016) developed a two-level network for organisations and professionals. Specific studies within industries include Araujo's diachronic analysis of 301 Swiss banking elites during the 20th century through manual work (Davoine, Araujo, and Donzé, 2022), and his study on the transnational corporate elite network, in which he identified three career paths (Araujo, 2018). These studies collect data either from an existing database (Araujo, 2018; Davoine, Araujo, and Donzé, 2022), manually (Studer Matthias and Ritschard Gilbert, 2014), or do not disclose the collection process (Bühlmann et al., 2022).

In the data preprocessing stage, since job titles from resumes are typically human-written and unstructured, normalisation is necessary for effective data analysis. Common methods include semantic analysis using NLP techniques and the segmentation of titles or extraction of keywords. For instance, Zhang et al. (2019) utilised an online API, MonkeyLearn, to categorize original job titles into 26 classes based on resume descriptions, although their analysis primarily maintained domain-specific information. Extending this, Singh et al. (2023) included both domain and position information, while Zhang et al. (2021) extracted function and responsibility from titles, aggregating them by keywords based on the manually annotated IPOD dataset (Liu et al., 2019).

Sequence analysis has become an increasingly popular method in the social sciences (Abbott, 1983), but not many studies have been carried out on the careers of corporate executives. Henriksen and Seabrooke (2016) conducted sequence analysis to examine longitudinal corporate careers over different dimensions. Koch et al. (2017) studied status, employer, and function. CEOs holding multiple positions in a year were coded according to their highest-status role. However, this approach relies on manual labelling of statuses and excludes non-executive positions. Our research addresses these limitations by utilising automatic rank labelling and information reconstruction during database reconstruction.

3. DATA COLLECTION

Top managers are people who belong to the top arenas of these big corporations. The two top arenas of joint-stock companies are the Board of Directors and the Executive Committee. Therefore, in our paper, we refer to members of the Board of Directors and Executive Committees as top managers. In many Anglo-Saxon companies, the Executive Committee does not formally exist; however, it is often constituted by executive directors. The expression "Executive Committee" tends to be generalised, nevertheless, some firms have their own appellation: "Executive management", "Management Team", "Management Committee" and so on. While in Germany and German-speaking countries, the two arenas are the Supervisory

Board and the management board. Despite the variation in terminology, these structures all refer to the same concept. They are leading the company on the executive, and top management levels. Executive Committee arenas include chief officers: Chief Executive Officer, Chief Financial Officer, Chief Operating Officer, and so on.

Biggest firms are joint-stock companies that belong to national main stock exchange indices: Dow Jones, Nasdaq (USA), FTSE 100 (UK), CAC 40 (France), SSE (China), and so on. Since we are studying the career trajectories of executives in large corporations based in France, our data source is CAC 40 Company executives. (CAC is the market index to track the stock of 40 companies with the highest capitalisation value.) Therefore, we identify members of Board of Directors and Executive Committees from those large corporations according to their financial report of 2019. We then retrieve their profiles from LinkedIn, as this seems to be one of the most popular and most frequently used websites for public resume information. This approach aligns with other job mobility-related literature (Li et al., 2017; Singh et al., 2023; Vafa et al., 2022; Zhang et al., 2021).

In our paper, the acquisition of CVs plays a pivotal role in constructing a comprehensive and informative database. First, we identify from the annual report all members (n=845) of the Board of Directors and Managing Committee as of 31st December 2019 of the 40 firms belonging to the French stock market index CAC 40. Then we use PhantomBuster to collect LinkedIn URLs based on name and institution information. After that, we choose to use iScraper API, a third-party service specializing in collecting LinkedIn profile data, as our primary data source. The choice to utilise API over web scraping was driven by a comprehensive evaluation of legal, ethical, technical, and operational factors. LinkedIn's terms of service restrict unauthorised scraping activities. In contrast, an API provides a tailored solution to structured, accurate data, significantly reducing the time and resources required for data preprocessing. We choose the JSON (JavaScript Object Notation) format to receive data considering its widespread acceptance as a standard for data interchange. JSON's text-based nature and language independence make it highly human-readable and machine-friendly. The lightweight nature of JSON minimises bandwidth usage, making it an ideal format for transmitting data over the internet. Of the 845 individuals, we successfully collected 363 usable CVs on LinkedIn.

4. RELATIONAL DATABASE CREATION

Our database consists of 11 tables, each designed to store specific types of data to ensure that all information extracted from LinkedIn CVs is included and can support the analysis of career paths. Specifically, there are tables to store fixed information (e.g., locations, certificates, skills, etc.), but the main entities include: an individual, defined by surname, first name, sex, and date of birth; an institution, defined by name, type (either professional or educational), nature (e.g., firm, public administration, NGO, etc.), scope levels (e.g., group level, area level, country level, division level), and industry sectors; a professional position (e.g., CEO, Chairman, executive vice-president, head of, etc.); an arena (e.g., board of directors, executive committee, etc.); and a trajectory, which may be professional or educational, which links these different entities through recorded time intervals. Each table uses primary keys (e.g., "id_individual" in the individual table) to uniquely identify records, while foreign keys establish relationships between tables, creating a coherent relational structure. For instance, the education table includes a foreign key "id_institution", referencing the primary key in the institution table.

Similarly, the career table contains foreign keys “id_individual”, and “id_institution”, linking it to both the individual and institution tables to represent specific stages in a person's career. The career table also connects to the position–title–arena table through its primary key, allowing for the inclusion of additional contextual information about each role. The individual table is linked via foreign keys to six other tables: career, education, industry, location, certificate, and skill, enabling the integration of diverse information about each individual.

The schema of our relational database can be found in Figure 2. It is derived from the structure of LinkedIn resumes, ensuring that key elements commonly found in LinkedIn profiles are fully represented. However, the schema has been designed with flexibility in mind and can be adapted to accommodate data from other sources. Apart from original information from LinkedIn CVs, we also enrich this database in the following aspects.

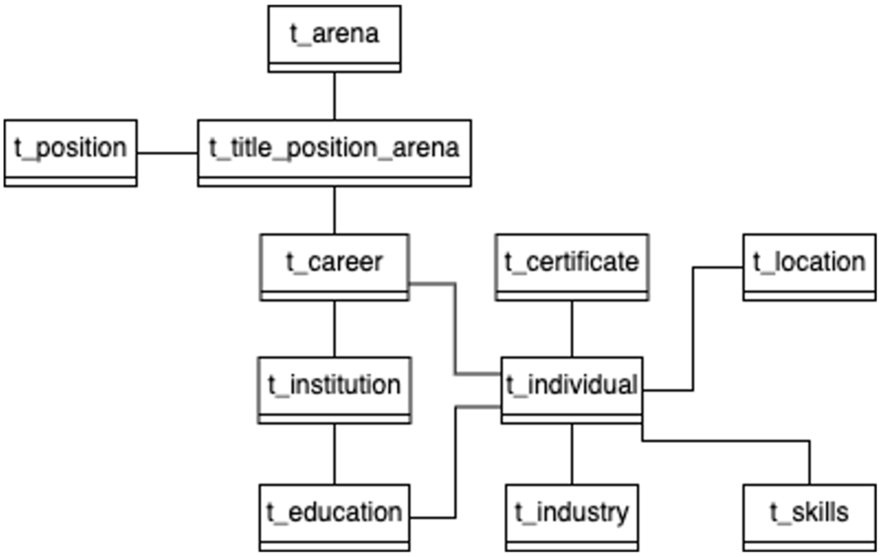

Figure 2. Relationship database schema

4.1 Career

In our database, a career trajectory is represented as a sequence of an individual’s jobs, institutions, and locations over specific time intervals (e.g., start year, end year), which are stored in the career table. To support further analysis, we also include a "career step" indicator that captures the specific stage within the individual’s professional trajectory. Since professionals may hold multiple positions simultaneously, we use the position–title–arena table to store detailed information for each occupation (e.g., arena, position, rank, etc.). We describe these components in more detail in the following sections.

4.2 Arena

An arena is an established social milieu where people physically meet on a regular basis, usually for the purpose of coordinating the activities of an organization. It is typically a formally collegial space where participants are, in principle, allowed to speak and express their views. By extension, it is also a place of debate where confrontations can arise. Boards of directors, executive committees, commissions, ministerial cabinets, etc., are considered arenas. Not all positions performed by an individual take place in or are directly attached to an arena. Examples include CEO, Manager, Professor, and Engineer. In this case, we use a default arena, which is "Executive position". This variable is also useful when individuals take positions in both executive and non-executive roles, allowing people's trajectories to be classified into main trajectories (executive positions) and secondary trajectories (other arenas).

4.3 Job Title Reconstruction

The "Job title" information we collect from resumes is often unclear, complicating subsequent analysis. This item might include information involving multiple aspects: job responsibilities (e.g., human resources, finance, commercial, operating), positions (manager, director, head), arenas (board of directors, strategic committee, management committee), and extra information (for example, if the position is within a subsidiary). Normalising job titles is crucial to address challenges such as the following: the vast diversity of over 1000 different positions without normalising, language discrepancies (e.g., "Director General" vs. "CEO"; varying responsibilities under the same title (e.g., a chief officer might manage finance or technology), and inconsistent naming (e.g., "Global CEO" vs. "CEO"); To tackle these issues, we categorise the raw data from LinkedIn into three attributes as shown in Figure 3: "position", "arena", and "title description" where "title description" captures additional details beyond standardised "position" and "arena" categories. This structured approach helps in accurately interpreting and grouping job titles for analysis.

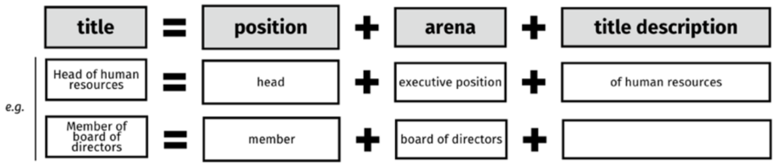

Figure 3. Title_reconstruction

4.4 Rank

One significant variable we include in this table is rank, which is used to show professionals' progression in their careers. We classify all positions into executive and non-executive ones. There are 10 ranks for executive positions (i.e., 1 - Entry-level or specialist, 2 - Middle-level management, 3 - Senior management, 4 - Executive committee, 5 - Deputy CEO, 6 - CEO, 7 - Chairman and CEO, 8 - Founder, 9 - Job outside of big corporations, 10 - High civil servant) and 7 ranks for non-executive positions (i.e., a - Chairman of the board of directors, b - Vice chairman, independent leader director of the board, c - Chairman of committee of the board of directors (or supervisory board), d - Member of committee of the board of directors (or supervisory board), e - Member of the board of directors, f - Non-executive member of the board of directors, g – Advisory arena).

4.5 Education

In addition to LinkedIn data, we also search for public records for individuals who do not have educational information listed. There are currently 823 records in Table t-education. Based on the values in the "degree name" and "field of study" columns, we categorize these into different types. We would like to note that education experiences without specific descriptions under "Grande École" are labelled as "Master's Degree and Equivalent" In the executive education category, we include the Executive MBA and similar programs, while the standard MBA is classified as "Master's Degree and Equivalent".

4.6 Institution

Our database includes over 1,000 institutions, which we categorise by nature, type, scope level, and parent institution when applicable. There are two main types: educational and professional. Nature is divided into five categories: non-profit, corporations, academic institutions, government entities, and others. We initially employ a heuristic-based Python script using specific keywords to classify these institutions—for example, "inc", "ltd", "corporation", and "GmbH" for corporations; "ministry", "department", and "federation" for government entities; "foundation", "institute", "society", and "association" for non-profits. Institutions marked as "others" are re-evaluated using already labelled institutions as new keywords. For further refinement, we utilise the latest capabilities of ChatGPT (as of April 2023) to assist in categorisation. However, some institutions remain uncategorised due to insufficient details. Regarding scope, corporations are classified into four levels: group, area, country, and division, with "area level" encompassing multiple countries or large countries like the US. ChatGPT also supports this aspect of our categorisation effort.

5. AUTOMATIC RANK LABELLING

Rank labelling is a classification task. As we are incorporating the career hierarchy variable "rank", it is necessary to label this feature for all career trajectories based on the 10 executive ranks and 7 non-executive ranks. Traditionally, this task has been carried out manually. However, in our paper, we aim to reduce the amount of manual work and automate the process to the maximum extent possible.

5.1 Method

We employ a combination of rule-based and machine-learning methods to address the rank labelling task. Specifically, the rule-based method is suitable for straightforward situations in rank labelling, while machine-learning methods are better suited for more complex scenarios.

Machine learning is particularly applicable due to the presence of textual data ("title description") and ordered data ("career step"). For example, the job title "CEO American" and "CEO" in a subsidiary may not denote a CEO in the true sense. Therefore, we use machine learning methods to classify such complex situations when executive ranks are at levels 1, 2, or 3. As mentioned in previous sections, the criteria for rule-based methods are clear. Thus, this section focuses on selecting variables in complex situations using machine learning techniques. There are four factors to consider: the title description provides additional information for job titles; positions; scope levels, and the specific step of this trajectory in the professional's career. Therefore, the four relevant variables are "id position", "scope level", "title description", and "career step".

5.2 Use of Classification Models

The target variables in this classification task are Ranks 1, 2, and 3. We have considered five models to build the classifiers: (a) K-Nearest Neighbours (KNN) is a non-parametric, lazy learning algorithm that predicts the class of a data point based on the majority class among its k nearest neighbours, suitable for irregular decision boundaries. (b) Support Vector Machine (SVM) with an RBF kernel optimally separates classes in high-dimensional spaces by maximizing the margin between them, which is ideal when dimensions outnumber samples. (c) XGBoost, a fast and efficient implementation of gradient boosted decision trees, offers flexibility across various predictive modelling problems. (d) Decision Trees provide a simple, interpretable tree-like graph of decisions, though prone to overfitting. (e) Random Forest, an ensemble method using multiple decision trees, excels in accuracy for both categorical and continuous data.

5.3 Implementation

To train this classifier and predict other unlabelled records using this classifier, we follow the workflow outlined in Figure 4. Step 1: Data preprocessing. Textual Data ("title description") is converted into numeric vectors using Term Frequency-Inverse Document Frequency (TF-IDF) vectorisation. Categorical features ("id position", "scope level") are transformed into binary matrices by one-hot encoding. Ordered data ("career step") is also standardised. Additionally, given the underrepresentation in our training datasets—Class 1 (56 records) and Class 2 (140 records) compared to Class 3 (479 records)—we employ SMOTE (Synthetic Minority Over-sampling Technique) to adjust class weights. Step 2: Training and 10-fold cross-validation. We utilise 676 labelled career trajectories to build the classifier. The SVM model with an RBF kernel demonstrates the best performance among all, achieving an average accuracy of 86%, as depicted in Figure 5. For the SVM method, we generate an overall confusion matrix for all the cross-validation folds. This is accomplished by employing "cross val predict" to obtain predictions for each instance while it is in the test set and then calculating the confusion matrix for these predictions against the actual labels. According to confusion matrix results (Figure 6), 42 instances (75.0%) in Class 1, 102 instances (73%) in Class 2, and 438 instances (91%) in Class 3 are predicted correctly. Class 3 exhibits the highest number of correct predictions, indicating robust classifier performance for this class. Before employing the classifier to predict ranks for unlabelled records, we perform Step 3 to ensure the accuracy. In Step 3, we identify 100 mismatched records by comparing actual and predicted ranks during the cross-validation process. Using the k-means method, evaluated by the elbow method and silhouette scores, we select the number of components as 20, with a cumulative explained variance of 0.7. We identify two clusters (Figure 7): Cluster 0 contains 68 records, and Cluster 1 contains 32 records. These clusters have a commonality; the scope level is at the group level, but the actual ranks for more advanced career steps are higher in Cluster 0, and those for earlier to mid-stage career steps are lower in Cluster 1. Step 4: We extract 344 similar records from the 1370 unlabelled records by calculating their distance. These selected records are likely mislabelled by machine learning classifiers; thus we update them manually, and the remaining 1026 records are predicted by the classifier directly.

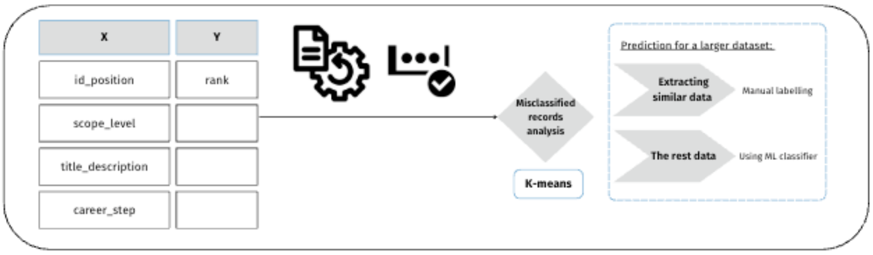

Figure 4. Rank labelling automation

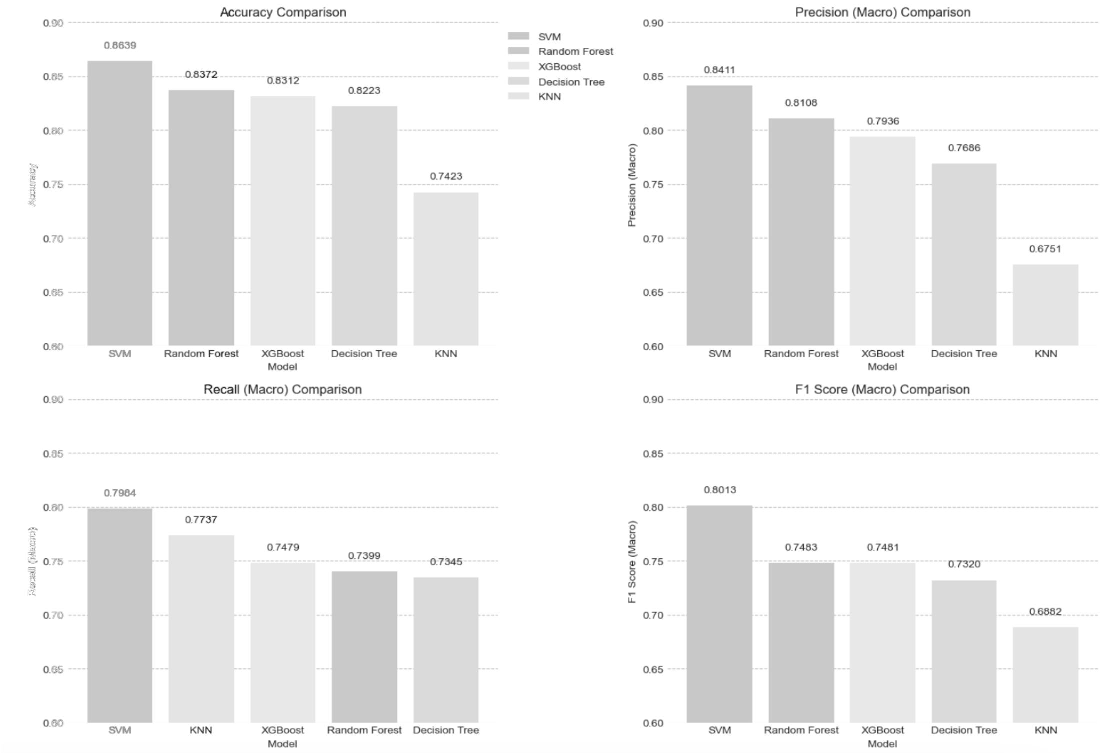

Figure 5. Classification model comparison

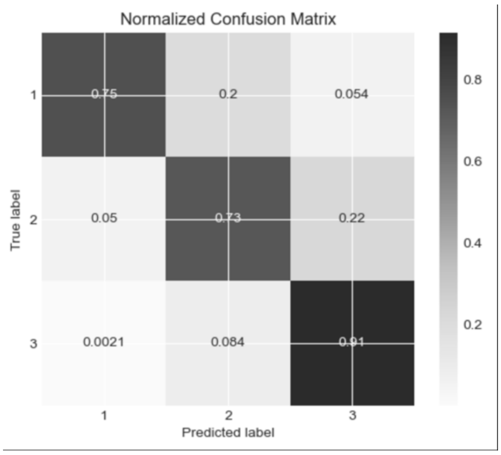

Figure 6. Confusion matrix

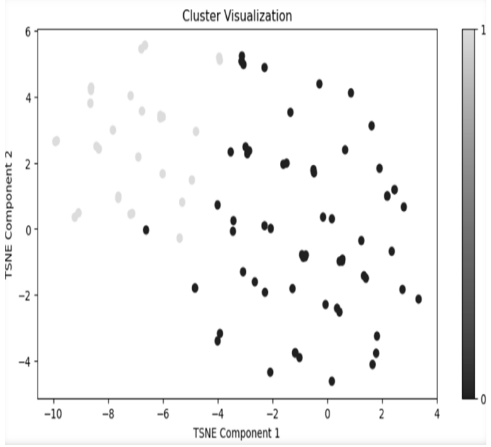

Figure 7. Clustering visualisation of misclassified samples

6. GRAPH DATABASE

Neo4j (Neo4j, Inc., n.d.) organises data in a graph format. Nodes represent entities such as people, products, or concepts, while edges (or relationships) connect these nodes, illustrating how they are related. This model is highly effective for representing complex, interconnected data. Unlike relational databases, which organise data into multiple tables, graph databases are schema-free and align more closely with natural human cognition. They excel in applications like recommendation engines, social media platforms, and knowledge graphs.

Given our existing relational database, which contains comprehensive information on the corporate careers of top managers, we prepare to create a graph database in Neo4j. We selectively scale down the data from our relational database, choosing only the most critical entities for our graph to maximize the benefits of the connections. Specifically, we include information from tables on individuals, education, careers, title-position-arena, and institutions. The basic principles of translating a relational data model to a graph data model are straightforward: a row becomes a node, a table name serves as a label name, and a join or foreign key represents a relationship. When modelling working trajectories, we draw inspiration from the work of Constantinov et al. (2020). On the one hand, we incorporate a hierarchical time model using "year" nodes to reduce computational time. On the other hand, we link consecutive positions with relationships called "next-career" to simplify the visualisation of an individual's career path. Specifically, our graph comprises seven nodes: Arena, Title-Position-Arena, Career, Year, Education, Individual, and Institution. Figures 8 and 9 are examples that show how career node and individual node connect with others. In the Future Work section, we will describe how we utilise the graph database for further analysis.

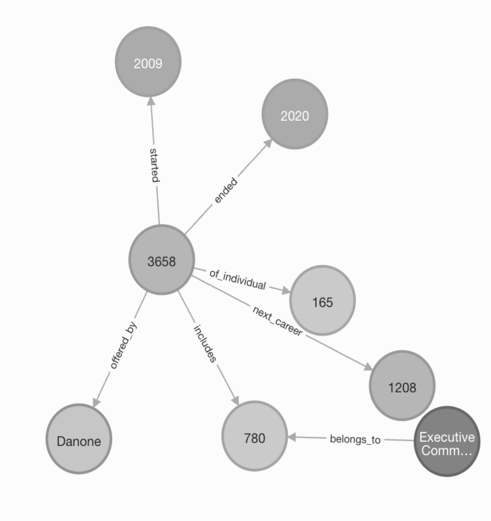

Figure 8. Node: career 3658

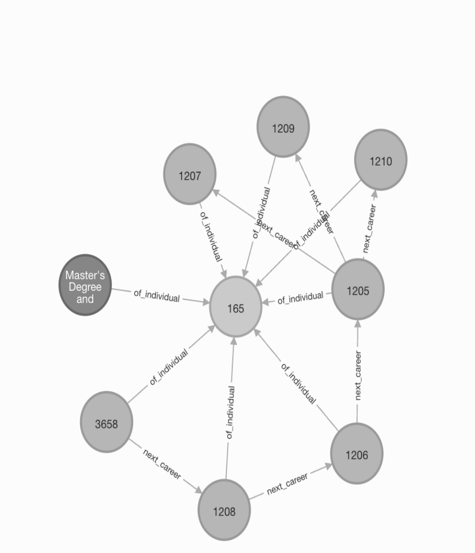

Figure 9. Node: individual 165

7. SEQUENCE ANALYSIS

Sequence analysis is used to explore social and behavioural phenomena like career paths and educational trajectories, focusing on how elements are temporally ordered. This method facilitates the examination of patterns and transitions across various industries and educational backgrounds. We also perform parallel analyses for executive and non-executive positions, acknowledging that top managers might occupy positions in both categories. For these analyses, we employ "TraMineR", an R-package designed for sequence data.

7.1 Data Preparation

While TraMineR supports various ways of organising sequence data, in this analysis, we opt for one of the most popular forms—the STS format. In this format, each sequence is presented as a (row) vector of consecutive states. Consequently, it is necessary to reshape the data first, which involves reorganising the data so that each individual's career states over time are laid out across columns. This method requires a continuous sequence of "career-year" values for each individual without overlapping or gaps. To address the issue of overlapping, we transform the records into along format by extending the start-year and the end-year into different career years. For the first scenario of overlapping issues, if the end-year of one record is the same as the start-year of the subsequent record, we maintain the subsequent one. For other scenarios of overlapping issues, we choose a higher rank among multiple career trajectories in the same career year. At this stage, we focus on the main trajectory analysis, so we set a "rank" map in Python. This mapping assigns numeric ranks from 1 to 7, where 7 represents the highest and 1 the lowest within this subgroup. Ranks 8, 9, and 10 are treated as equivalent and less significant than Ranks 1–7, though they remain higher than any alphabetic ranks. Alphabetic ranks are considered the lowest overall. This approach ensures that executive positions are assigned higher ranks for prioritization in the analysis.

7.2 Exploring Trajectories

We analyse career trajectories of 324 individuals, defining 11 distinct states across two dimensions: executive vs. non-executive roles, and positions within vs. outside large corporations. According to our results (Figure 10), 73.4% of individuals start their careers in executive roles within corporations, with 26.8% at entry-level or as specialists, 32.4% in middle management, and 14.2% in senior management. Additionally, 14.5% start in government roles, while 12.1% begin in jobs outside big corporations, as founders or others. Our analysis, focusing on career lengths up to 37 years (Figure 11), reveals a trend where top positions such as CEO and Chairman become more common over time, whereas lower positions decrease. This analysis shows two key trends: top positions (such as those on the executive committee, CEO, and Chairman) are more prevalent, while lower positions (such as middle-level and entry-level employees) tend to diminish as the career year progresses.

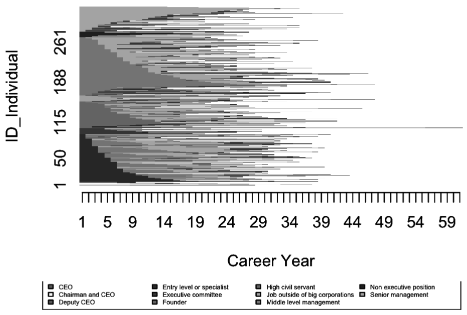

Figure 10. Index plot

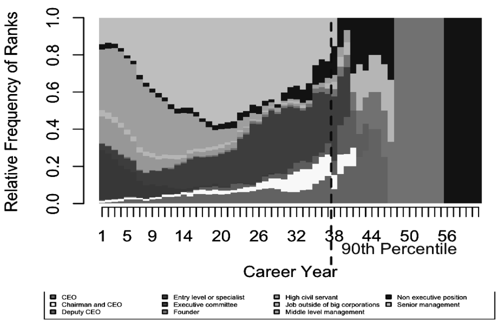

Figure 11. Rank distribution

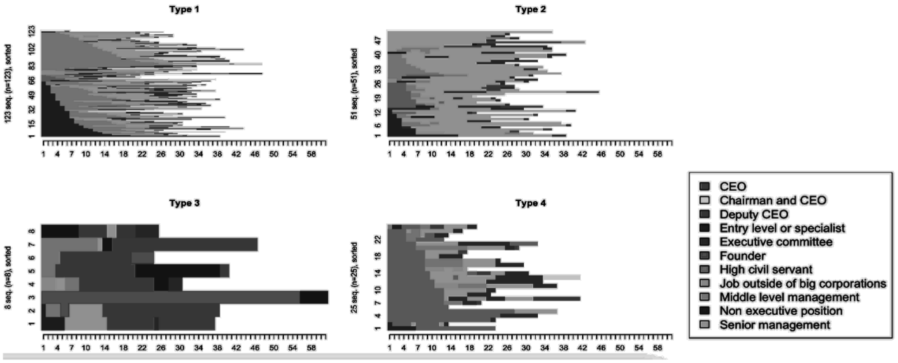

Figure 12. Optimal matching for people whose career information starting before 30

7.3 Typology Analysis

Optimal matching measures pairwise dissimilarities between sequences and then identifies types of patterns by analysing the sequences based on these dissimilarities (Studer Matthias and Ritschard Gilbert, 2014). We conduct optimal matching for individuals in our database, excluding individuals whose career information starts too late (after age 30) to ensure accuracy, as the majority of these sequences start from senior management, executive committees, non-executive positions, Chairman, and CEO roles. In total, 207 people are involved in the typology analysis. The clustering as shown in Figure 12 reveals four types of paths. Type 1 brings together long careers, with individuals starting out more often than elsewhere in the lowest positions and staying relatively long in middle management jobs before reaching executive positions after 15-20 years. Type 2 is a fast-paced career in which individuals, even if they start out in a junior managerial position, quickly rise to the top (around 10 years, sometimes less). Type 3 includes individuals who have been CEOs for most of their career. They are very few in number and would benefit from a more qualitative analysis to understand exactly what they represent. Type 4 includes executives who began their careers in the French civil service. They embody a well-known phenomenon in France known as "pantouflage", when a senior civil servant is hired by a company. It is interesting to compare Types 1 and 2, as they present the same career structures but at different paces. The percentage of women is lower in Type 2 (21.1%) compared to Type 1 (31.7%). There are also fewer non- French individuals in Type 2 (7.81%) than in Type 1 (11.38%). Regarding education, 76% in Type 2 attended Grande École, slightly higher than the 73% in Type 1; however, the proportion attending the most prestigious institutions such as École Polytechnique, HEC (École des Hautes Études Commerciales), and ENA is much higher in Type 2 (37.3%).) than in Type 1 (24.4%)

8. CONCLUSION AND FUTURE WORK

In this paper, we have introduced our workflow for enabling data analysis of corporate careers, encompassing three data stores: the CV store, the relational database store, and the graph database. This ready-to-use application is poised for future applications in data analysis within the corporate career field. Additionally, we have incorporated sequence analysis as an exemplar of future analytical approaches.

For our future work, we are exploring further analyses of our databases. Specifically, we are refining typology analysis in sequence analysis to uncover different career types and the reasons behind them in our database. We are also working on network analysis with our graph database. For instance, temporal visualization of institutions and individuals can provide insights into how connections evolve over time. Concurrently, we are investigating career hubs to explore the trajectories individuals take before arriving at CAC 40 firms.

ACKNOWLEDGEMENT

Parts of this research were carried within the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 945304 - Cofund AI4theSciences hosted by PSL University.

REFERENCES

Abbott, A. (1983). Sequences of Social Events: Concepts and Methods for the Analysis of Order in Social Processes. *Historical Methods: A Journal of Quantitative and Interdisciplinary History,* Vol. 16, No.4, pp. 129-147.

Araujo, P. (2018). Dynamics of internationalization: a sequential analysis of the careers of Swiss banking elites. In O. Korsnes et al. (eds.) *New directions in elite studies*. Routledge, pp. 73-89. https://doi.org/10.4324/9781315163796-4

Bühlmann, F. et al. (n.d). How career hubs shape the global corporate elite. *Global Networks* n/a. https://doi.org/10.1111/glob.12430

Constantinov, C., Dogaru, D. and Mocanu, M. (2020). Graph Model Proposals for Capturing Meta-information Within Professional Network Data. *2020 Seventh International Conference on Social Networks Analysis, Management and Security (SNAMS)*. IEEE.

Davoine, E., Araujo, P. and Donzé, P.-Y. (2022). Continuity and discontinuity in the capitals legitimating business elites – a diachronic analysis of the Swiss banking elites during the 20th century (1890-2020). *Conference of the European Group of Organization Studies (EGOS)*, Calgiari (Italy), online.

Davoine, E. and Schmid, S. (2022). Career patterns of top managers in Europe: Signs of further globalisation? *European Management Journal*, Vol. 40, pp. 467-474. https://doi.org/10.1016/j.emj.2022.05.007

Han, Y., Colazzo, D. and Dudouet, F.-X. (2025). A General Methodology Enabling Data Analysis of Corporation Careers. *IADIS Information Systems 2025.* Madeira Island, Portugal (To appear).

Henriksen, L. F. and Seabrooke, L. (2016). Transnational organizing: Issue professionals in environmental sustainability networks. *Organization*, Vol. 23, pp. 722-741. https://doi.org/10.1177/1350508415609140

Khan, S. R. (2012). The Sociology of Elites. *Annual Review of Sociology*, Vol. 38, pp. 361-377. https://doi.org/10.1146/annurev-soc-071811-145542

Koch, M., Forgues, B. and Monties, V. (2017). The Way to the Top: Career Patterns of Fortune 100 CEOS: The Way to the Top: Career Patterns of Fortune 100 CEOs. *Human Resource Management*, Vol. 56, pp. 267-285. https://doi.org/10.1002/hrm.21759

Koyuncu, B., Hamori, M. and Baruch, Y. (2017). Guest Editors' Introduction: CEOs' Careers: Emerging Trends and Future Directions. *Human Resource Management*, Vol. 56, pp. 195-203. https://doi.org/10.1002/hrm.21760

Li, L. et al. (2017). NEMO: Next Career Move Prediction with Contextual Embedding, *Proceedings of the 26th International Conference on World Wide Web Companion - WWW '17 Companion*.

Liu, J. et al. (2019). *IPOD: Corpus of 190,000 industrial occupations*. https://doi.org/10.48550/arXiv.1910.10495

Reeves, A., Friedman, S., Rahal, C. and Flemmen, M. (2017). The Decline and Persistence of the Old Boy: Private Schools and Elite Recruitment 1897 to 2016. *American Sociological Review*, Vol. 82, pp. 1139-1166. https://doi.org/10.1177/0003122417735742

Singh, S., Gupta, A., Baraheem, S. S. and Nguyen, T.V. (2023). Multi-Output Career Prediction: Dataset, Method, and Benchmark Suite. *2023 57th Annual Conference on Information Sciences and Systems (CISS)*.

Studer, M. and Ritschard, G. (2014). *A comparative review of sequence dissimilarity measures*. https://doi.org/10.12682/LIVES.2296-1658.2014.33

Vafa, K. et al. (2022). *CAREER: Transfer Learning for Economic Prediction of Labor Sequence Data*. Working Paper No. 4074. Available at: https://www.gsb.stanford.edu/faculty-research/working-papers/career-transfer-learning-economic-prediction-labor-sequence-data

Zhang, C., Wang, H. and Wu, Y. (2019). ResumeVis: A Visual Analytics System to Discover Semantic Information in Semi-structured Resume Data. *ACM Transactions on Intelligent Systems and Technology*, Vol. 10, pp. 1-25. https://doi.org/10.1145/3230707

Zhang, L. et al. (2021). Attentive Heterogeneous Graph Embedding for Job Mobility Prediction. *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*.