Structural Analysis of the Danish Job Market using Network Science

Elli Georgiou, Henrietta Domokos, Razvan Ciubotaru

This manuscript was compiled on December 6, 2023

Graph theory represents the backbone of social networks, traditionally used in interactions and connections among individuals. However, these networks may be used in analyzing any complex system, providing crucial insights about the system entities and the way that connections are created between them. In this project, we use advanced graph theory methods to perform an in-depth investigation of the job market landscape. We built a dynamic social graph to represent the complex relationships that exist within the job market system by utilizing a dataset that includes key features about job postings. Our research focuses on discovering common skill-demand patterns, identifying key players across industries, and understanding the dynamic structure of professional networks. Our goal is to uncover insights about job market trends, talent interdependencies, and potential career trajectories by using particular graph algorithms. The results of this research will likely improve understanding of the complex relationship between demand, skills, and employment prospects in today's job market, providing essential views for both industry professionals and academics.

Job market | Social networks | Graph Theory | Natural Language Processing

Significance Statement

By using data science methods we were able to analyze a large number of LinkedIn job descriptions. We detect patterns in skill demand, identify major industry trends, and explore the connections between different job types. This analysis offer us a better comprehension of the job market in today's digital world. Our project aims to provide important and helpful information for both job seekers and employers regarding in-demand skills and current employment trends. Essentially, it is like creating a road map of the job market, making it easier for people o navigate their career paths and for companies to find the right talent.

Different studies were conducted in the recent years in the job market analysis, all of them trying to enable job seekers to better understand the requirements and challenges that arise when looking for a job. However, few of the studies were conducted regarding the Danish job market. With an increasing number of students that are constantly looking to start or continue their professional career in Denmark and with a vast number of jobs to choose from, we conducted an in-depth investigation of the Danish job market while analysing the job offers available on the social platform LinkedIn.

Firstly, we explore in our research the dynamics of job hunting by analyzing a large dataset of job descriptions gathered through the LinkedIn API. To begin with, we extracted and carefully cleaned an important amount of jobs data in order to improve the quality of these descriptions. A key phase in our analysis was categorizing these descriptions based on their various industries, paving the path for extensive industry-specific assessments.

Furthermore we used the Term Frequency-Inverse Document Frequency (TF-IDF) method (1) for exploratory data analysis with the goal of figuring our the relevance and importance of specific words in the job market sector. A critical component of our research was also the creation of a comprehensive graph that included characteristics such as industry, firm and location. This graph effectively depicts the interrelationships between various employment positions and industries, with nodes representing individual job descriptions and edges indicating TF-IDF similarities.

Moving forward we did a detailed community analysis using the graph to uncover patterns and clusters in the job market. Additionally, we performed sentiment analysis on the job descriptions which gave us significant insights into the prevailing sentiments and perceptions across different industries.

By combining TF-IDF analysis, graph-based modeling and sentiment analysis, our project depicts a fresh look at the complexities of the employment market, revealing trends and linkages that would otherwise go unnoticed. The results and the findings of this project are extremely valuable to job searches, recruiters and industry analysts offering a detailed data driven view of the evolving job market.

Results

The data. There are many LinkedIn datasets available online which capture important job insights and details, such as (2). However, none of them refers to the Danish job market, that is why we created our own dataset. To gather data essential for

Equal contribution - Abstract; Introduction | Elli Georgiou - Significance Statement; Results: Job similarity, Clustering, Communities; Discussion | Henrietta Domokos - Results: Data, Job similarity, Sentiment analysis; Materials and Methods | Razvan Ciubotaru - Results: Data, Job similarity, Graph analysis; Discussion; Materials and Methods

Table 1. Example of a job sample in the proposed dataset

Feature	Value
1. Company	FUJIFILM Biotechnologies
2. Title	Microbiologist in quality control
3. Location	Capital Region
4. Number of applicants	104
5. Experience	Mid-Senior level
6. Employment type	Full-time
7. Industry	Healthcare and Pharmaceuticals
8. Job description	Are you eager to join a fast-growing team...

analyzing the present job market and developing our solution, we opted to extract information from publicly available job postings on LinkedIn directly. We developed a specialized web scraper capable of not only retrieving the job descriptions but also capturing additional metadata, including the job title, company name, number of applicants, industry categorization and other relevant details.

In order to clean the dataset, we eliminated the duplicate jobs, along with the non-English job descriptions. Jobs with no industry or location where also eliminated. This collection resulted in 971 samples, each one of them featuring 9 attributes. An example of a scraped job poster is presented in table 1. In this research, the most important feature of the dataset is the **job description**, since the graph is built based on it. Hence, pre-processing the these descriptions is crucial in order to determine the similarity between jobs.

Job similarity. The first method we used to compute the similarity is based on transformers, a strong Large Language Model (LLM) which uses the attention mechanism in order to determine the semantic importance of each word (3). A problem of the transformers is the limited number of input tokens, for example, BERT (4) accepts maximum 512 tokens. Truncating the job description tokens to this limit leads to information loss. To overcome this issue, we used a particular transformer, the *Longformer* (5), which accepts 4096 tokens. The model creates vector embeddings with 768 elements for each job description, and then cosine similarity is applied: $sim = cos\theta \in [0, 1]$, where θ is the angle between two vectors. Two job descriptions are more related as the similarity is closer to 1. However, the similarity values we obtained are $sim \in [0.93, 0.99]$, contrary to our expectations. We would have expected such high values for jobs in the same industry and lower values, below 0.7, for jobs in different industries. With such a narrow range, it was hard to find a threshold that would conclude the jobs similarity.

The second method is based on the TF-IDF measurement: for each job description, we created a dictionary, where each word had a significance score. The most significant words in the job market are presented in the word cloud in figure 1.

After ordering these words according to their score and selecting the top 15 words, we concluded that two job descriptions are similar if at least 5 words are common in this top.

Graph analysis. Each node of the graph is represented by the job description ID found in the dataset, and its attributes are: job description, company, location and industry. If two jobs are similar, an edge is drawn between the corresponding nodes. One topic of interest in this research is to observe the similarity of jobs in different industries. After creating the edges based on the aforementioned criteria and eliminating the nodes with no connections (159 nodes), the resulting graph featured **727 nodes** and **3202 edges**. A representation of the graph based on the job industries is displayed in figure 3. Out of 817 connections between jobs from different industries, most of them are made between jobs in the Financial, IT, Manufacturing and Services industry. One reason for this is because most of the jobs in the market industry are concentrated around these industries: IT - 170 jobs, Financial - 133 jobs, Manufacturing - 159 jobs and Services - 87 jobs. Moreover, with an increasing demand of digitization and automation in almost all industries, we find this aspect normal, since most industries need IT workers in order to fulfill these requirements. For example, in the US, the number of in-demand IT jobs in the financial sector increased by 28% in the first half of 2022 (6).

Fig. 1. Word cloud of the Danish job market. These words are naturally met in any job description, with emphasis on work-related terms. The top 5 words and their TF-IDF scores are: team (0.048), work (0.038), experience (0.037), project (0.036), business (0.034).

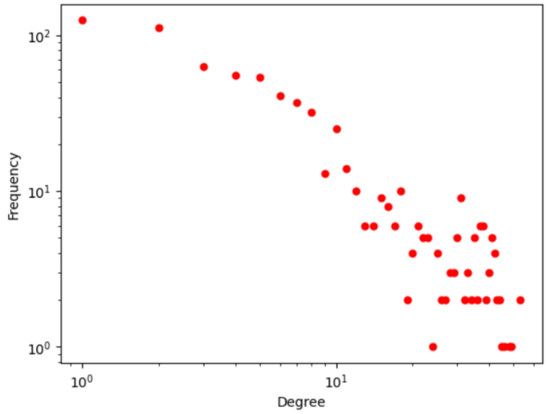

Fig. 2. Degree distribution - logaritmic scale. We considered this scale for illustrating that the values follow a power-law distribution, hence the graph is not considered a random network, but rather a real network.

The node degree distribution is shown in figure 2. The majority of the nodes have the lowest degree $k_{min} = 1$, and just one node has the highest degree, $k_{max} = 54$. An important aspect highlighted in this figure is the existence of hubs (nodes with high degree). These hubs only exist in real networks, and it states that high-degree nodes can coexist with small-degree nodes, and they may also lead to the formation of different communities. The degree exponent is

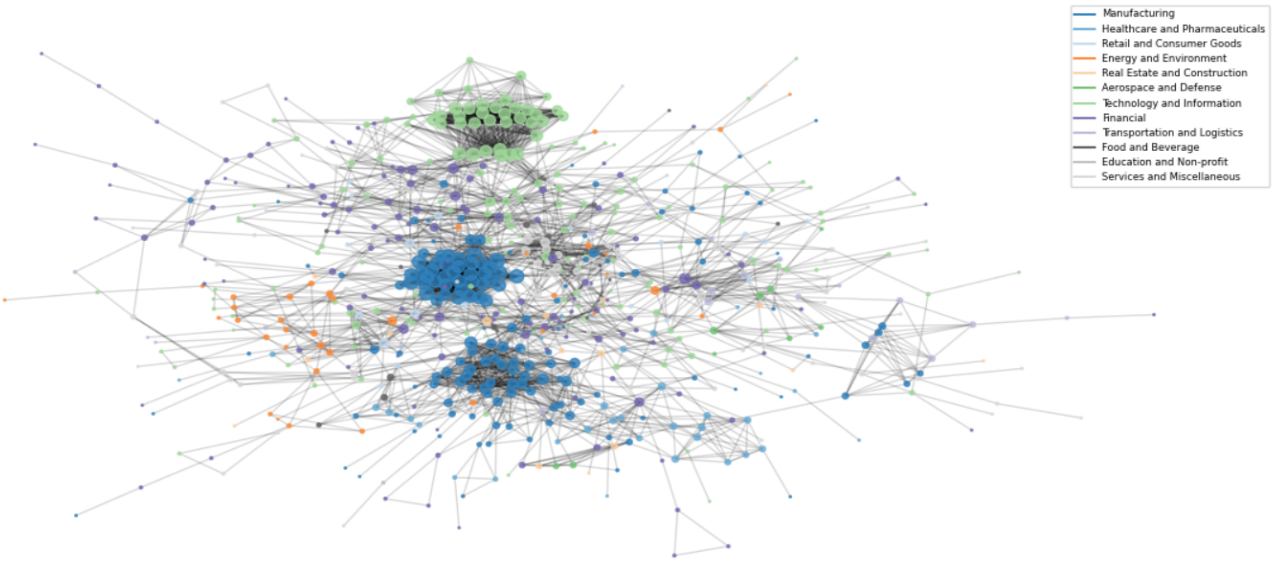
Manufacturing
Healthcare and Pharmaceuticals
Retail and Consumer Goods
Energy and Environment
Real Estate and Construction
Aerospace and Defense
Technology and Information
Financial
Transportation and Logistics
Food and Beverage
Education and Non-profit
Services and Miscellaneous

Fig. 3. Job market graph. Size of the nodes is proportional to their degree: the higher the degree, the bigger the node. It can be seen that mostly jobs in the same industry tend to have a connection. However, 817 edges out of 3200 connect jobs from different industries.

$\gamma = 2.18 \in [2, 3]$, stating that our graph is a scale-free network, which has the ultra-small world property: $\langle d \rangle \sim \ln(\ln(N))$, where $\langle d \rangle$ is the average shortest path between any 2 nodes, and N is the number of nodes.

When visually analysing the graph, it can be seen that nodes with a high degree tend to be plotted close to each other, forming communities. The nodes with the highest degree belong to the Manufacturing industry. After checking the top 30 nodes with respect to their degree, 28 of them were jobs from the LEGO Company, all of them in the aforementioned industry, hence the high similarity of the job descriptions. Regarding the number of jobs, the top 5 comapnies are: The Hub - 48, Novo Nordisk - 48, LEGO - 43, GatedTalent - 23, HRtechX - 20. With a high market value, Novo Nordisk and LEGO are some of the biggest and most known Danish companies, and have always contributed significantly to the Danish job market (7), and the others are recruitment companies, which facilitate the job search process through the LinkedIn platform.

Clustering. In order to evaluate the clustering capabilities of the graph and prove its non-random nature, we generated a random graph using the Erdos-Renyi model as a reference, with the probability of connectivity $p = 0.015$ and number of nodes $N = 732$. The average clustering coefficient of our graph was notably higher, with a value of $\langle C \rangle_G = 0.584$ compared to $\langle C \rangle_{ER} = 0.015$. This high contrast highlights the non-random nature of connections in our job market network. The high value of clustering coefficient means that if a job is connected to two other job postings, there is a high probability that those two postings are also connected to each other. Such clustering suggests that jobs tend to cluster by industry or skills, which makes it easier for job seekers and recruiters to find what they are looking for. Furthermore, it emphasizes the importance of networking, since linked posts are likely to represent opportunities where industry connections play a major role in hiring.

Communities. To identify communities within our graph, we used the Louvain algorithm (8), which focuses on optimizing modularity. We iterated over an array of resolution values, from 0.5 to 3 with step of 0.1, to find the optimal resolution in an effort to maximize the modularity of our graph. The resolution parameter affects the size of community detection, more specifically, lower values lead to bigger communities, while higher values means smaller communities. Louvain algorithm detected 70 unique communities at the optimal resolution $\gamma = 1.3$ showing an important modular structure inside the network. With this set up we find that the highest modularity score is $M = 0.76$, which means that the network is strongly divided into clusters of job descriptions with similar characteristics, with 28 communities having all jobs in the same industry. The highest community, with 95 nodes, has most of the jobs in the IT - 23 nodes, Services - 23 nodes and Manufacturing - 23 nodes, which highlights again the high connectivity between these industries.

We compared this partitioning with a industry-level partitioning, considering each community being formed by nodes of the same industry. This resulted in 12 communities, with the modularity $M = 0.47$. Since the value is smaller, this would be *suboptimal partitioning*: it can be seen in the graph representation that some nodes in one industry are positioned among nodes in different industries, having a stronger connectivity to them.

Sentiment analysis. Nowadays, recruitment methods are evolving very quickly and companies are using newer and newer techniques in order to attract the most talented applicants. That's why it might be interesting to integrate the sentiment analysis into the assessment of job descriptions. In this context, sentiment analysis provides a nuanced understanding of the emotional impact that the language used by the recruiters may have on potential candidates.

The labMT 1.0 dataset (9), which is a selection of 10,222 words, was presented to Mechanical Turk users, who were

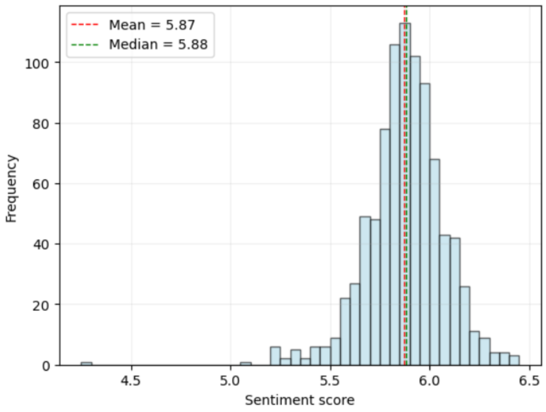

Fig. 4. Sentiment distribution histogram. With a broad range of sentiments $s \in [4.25, 6.44]$, there are certain patterns that can differentiate jobs and companies with respect to their sentiment score. The close values of mean: $\mu_s = 5.87$ and $median = 5.88$ suggest that the distribution is close to a symmetrical one.

tasked with assessing the happiness associated with each word. Words were then ranked based on their average happiness scores, resulting in a list ordered from the most positively perceived to the least.

The analysis assigns a sentiment score s to each job description, providing a quantitative measure of the overall sentiment. The histogram of the sentiment distribution is visualised in 4. However, since most sentiments are in the range $[4.25, 6.44]$, these values still point towards a rather neutral sentiment, since job descriptions tend to be written in an objective language, with little emphasis on subjectivity. Two job postings stand out in this analysis — one with the lowest sentiment score, which can be treated as an ”outlier” in the histogram, and one with the highest sentiment score.

In the case of the ”saddest job” at the company Paymentology, the sentiment score is $s = 4.25$. The corresponding job description reveals a focus on global fraud prevention, a domain that often involves the use of serious and cautionary language. The sentiment score reflects the nature of the responsibilities outlined in the job description. In contrast, the ”happiest job” at LEGO Group, with an average sentiment $\bar{s} = 6.3$, boasts a sentiment score $s = 6.44$. The job description emphasizes creativity, innovation, and a commitment to creating a positive and inclusive working environment. The score aligns with the company’s values of fostering a culture of play and imagination.

The analysis extends beyond individual jobs and companies, exploring the average sentiment of each industry. This provides a comprehensive view of how different sectors express the tone in their job descriptions. However, we found that the mean sentiment has a narrow range: $\bar{s} \in [5.81, 5.95]$, with the ”happiest” industry - Food and Beverages, and the ”saddest” one - Real Estate and Construction. This aspect ensures uniformity in terms of the language used in job descriptions, regardless of the industry.

Discussion

Looking on our project, there are some opportunities for improvement that we could have been explored with more time and data. We had a limit of pulling only 1000 job descriptions from LinkedIn per session and that was key restriction we faced. This limitation affected the depth and the variety of our dataset. With a bigger amount of data, our analysis probably would have included a wider spectrum of industries and employment categories.

The computational power required by LLMs was another limitation. It took 12 hours to compute the similarities between 500 jobs, and testing new models to find the right one for our research would take a lot of time, given the computational necessities.

Materials and Methods

Data scraping and processing. In order to acquire the dataset, we used scraping methods based on the `BeautifulSoup` library, where we looked for specific HTML tags in each jobs webpage. After collecting the data, it underwent an extensive preprocessing stage to enhance its suitability for subsequent analytical methods. Various text preprocessing techniques were applied to refine the job descriptions and conduct a comprehensive cleanup of other attributes. The preprocessing steps included:

- Word Filtering: Eliminating words containing numbers and special characters to enhance text clarity.
- Word Separation: Addressing concatenation issues, especially evident in words at the end of lines on LinkedIn (e.g., ”requirementsYou’re”), by separating them.
- Text Processing: Lowercasing, punctuation removal, tokenization, stop words removal, and lemmatization were performed to extract meaningful text.

Clustering coefficient. The clustering coefficient captures the degree to which the neighbors of a given node link to each other. For a node i with degree k_i the local clustering coefficient is defined as: $C_i = \frac{2L_i}{k_i(k_i-1)}$, where L_i represents the number of links from node i to its neighbours. The average clustering coefficient of a network is defined as: $\langle C \rangle = \frac{1}{N}\sum_{i=1}^{N} C_i$, where N is the number of nodes.

Modularity. It refers to the possibility of a graph network to be divided into separate and well-defined subgraphs called communities. The intuition behind modularity is to compare the actual wiring diagram within the network to what we would expect in a random network: $M_c = \frac{1}{2L}\sum_{i,j \in C_c}(A_{i,j} - p_{i,j})$, where L is the number of edges, C_c is a presumed community within the network,$A_{i,j}$ is the actual number of edges between nodes i and j, $p_{i,j} = \frac{k_i k_j}{2L}$ is the expected number of edges between node i and j in the randomized original network.

References

1. S Robertson, Understanding inverse document frequency: On theoretical arguments for idf. *J. Documentation - J DOC* **60**, 503–520 (2004).
2. Linkedin india dataset (https://www.kaggle.com/datasets/shashankshukla123123/linkedin-job-data) (2023) Accessed: 05-12-2023.
3. A Vaswani, et al., Attention is all you need. *CoRR* **abs/1706.03762** (2017).
4. J Devlin, MW Chang, K Lee, K Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding (2019).
5. I Beltagy, ME Peters, A Cohan, Longformer: The long-document transformer (2020).
6. High it jobs demand in finance (https://www.cio.com/article/405651/the-10-most-in-demand-it-jobs-in-finance.html) (2022) Accesed: 05-12-2023.
7. Novo nordisk slår lego af pinden i image-måling (https://www.berlingske.dk/business/novo-nordisk-slaar-lego-af-pinden-i-image-maaling) (2022) Accesed: 05-12-2023.
8. VD Blondel, JL Guillaume, R Lambiotte, E Lefebvre, Fast unfolding of communities in large networks. *J. Stat. Mech. Theory Exp.* **2008**, P10008 (2008).
9. PS Dodds, KD Harris, IM Kloumann, CA Bliss, CM Danforth, Temporal patterns of happiness and information in a global social network: Hedonometrics and twitter. *PLoS ONE* **6**, e26752 (2011).