Graph analysis of LinkedIn data-set

Ahmed Safwat, Ahmed Hamdy, Ahmed Fathi

May 24, 2023

Abstract

The purpose of this research is to provide useful information for jobs recommendation system on LinkedIn using graph analysis. By providing a data-set, a graph was constructed and analyzed using igraph and networkx libraries in Python. The degree distribution analysis is used to indicate the presence of companies within some parts of the graph. The connected components analysis also does indicate the relationships between the subgraphs and find the separated nodes that may be added to the larger components by recommendation. And there is also an important analysis which is community discovery that is the most significant analysis as it discovers the communities within the graph and that can be used to understand the replacements that may be done and recommend it to the employees within the community. And there are more analysis that will be discussed in the research later.

1 Introduction

The job market in Australia is a vast ecosystem where companies and employees interact, shaping the dynamics of various industries. Understanding the connections and dependencies between companies and employees can greatly impact decision-making processes, job search strategies, and industry research. In this paper, we delve into the extensive dataset we collected, focusing on the relationships between companies and employees in Australia. By representing these relationships in an unweighted undirected graph, we analyze its structure and properties, leveraging graph theory techniques to extract valuable information.

Utilizing the collected data, we constructed an unweighted undirected graph as shown in fig(1) that served as a visual representation of the connections between companies and employees. Each node in the graph represented either a company or an employee, while the edges represented the relationships or previous employment connections between them. This graph allowed us to observe and analyze the intricate web of connections within the Australian job market.

we present a comprehensive analysis of the network of companies and employees in Australia, employing a series of analysis techniques to gain deeper insights into their interrelationships and dynamics within the job market. Through

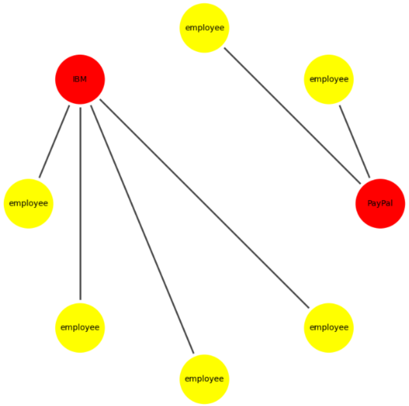

Figure 1: Graph Type

degree distribution analysis, we explore the overall connectivity of companies and employees. Path analysis uncovers potential career trajectories and pathways, while centrality analysis identifies key entities that play significant roles in the network.

Additionally, connected components analysis reveals clusters of closely connected companies and employees, unveiling communities and sub-networks within the job market. Clustering analysis groups similar entities based on attributes such as industry sector or geographical location, providing a nuanced understanding of specific industries or regions. Finally, community discovery analysis uncovers cohesive groups exhibiting shared connections and behaviors, identifying distinct subgroups within the job market.

By employing these analysis techniques, our study aims to offer comprehensive insights into the network of companies and employees in Australia. This knowledge contributes to efficient job matching and provides valuable industry observations, benefiting job seekers, recruiters, and policymakers in making informed decisions, understanding industry dynamics, and matching employees with the most suitable companies.

2 Methodology

The first important step in the project was to get a sample data to work on. After searching, we got to a pre-collected data from 2018 for about 10,000 people on LinkedIn as well as the jobs, companies they work at and other information. The data was on Kaggle website, it was enough, accurate and relatable to our goal. Using this data we can discover communities of employment and recommend jobs for the users. Unfortunately, there was a problem with the data as it contained a lot of information that is not useful for us. So, we firstly started filtration on Microsoft Excel to keep only the important data. we kept the only used 2 columns: memberUrn (unique number for each member) and companyName (the company he works at). Then the updated file was saved as csv file.

Using python library, csv, the data is read into reader object. Then it was extracted into 2 lists, edges and nodes. To construct the graph and start the analysis, the python igraph and networkx libraries were used. We added both the memberUrns and companyNames of each raw of the data as nodes to the graph. We checked the duplicates while adding the nodes and removed any repeated nodes. So, we can differentiate between the employees and companies by the name of the node, Employees are numbers and companies are strings. By passing these nodes and edges to the igraph library, the graph was constructed. After constructing the graph, the total nodes was 20610 and the number of edges was 29357. Then we started the analysis part and you can find below more details about the used algorithms and methods to complete the analysis.

3 Results and Analysis

3.1 Degree distribution analysis

This property is simply a basic concept in the business of networks and graph theory. It checks the distribution of degrees of nodes (connection numbers) in a given network. Also, it examines how the given connections are distributed across the network. In this graph, the bar chart shows a right-skewed distribution; also it shows that degree 1 is the most frequent with more than 10,000 nodes and the second most frequent is degree 2.

3.2 Path analysis

The path analysis between any two given in a graph studies the routes or paths that connect these two nodes and then analyzes the common relationships that these nodes have. It also counts the number of nodes in the route between the chosen nodes It helps analysts understand how information, influence, or other attributes flow between the two nodes. In this graph, the written code calculates all possible paths between the selected nodes then chose the shortest path and returns its length.

3.3 Centrality analysis (degree and betweenness)

Centrality is the measure of how important or prominent a node is in a network. It also helps in identifying more central or influential nodes in communication, information flow, or control networks; it has some different measures such as degree Centrality that classify the node as important if it has a higher degree (connected nodes) in addition to betweenness centrality that quantifies the extent to which a node lies on the shortest paths between other pairs of nodes.

In this graph, the analysis shows that some nodes are central or important because of the multi connections they have for degree centrality and some nodes are critical because of their existence in a huge number of shortest paths for betweenness centrality.

3.4 Connected components analysis

Connected components analysis in graphs is a technique used to identify groups of nodes that are interconnected and form distinct subgraphs within a larger graph. In a graph, a connected component is a set of nodes where there is a path between any two nodes in the set. So, the components analysis gives the unique subgraphs that are not connected with each other. This analysis is useful as it helps in understanding the overall structure and connectivity of a graph. By identifying connected components, we can identify distinct clusters or communities of nodes that share similar properties or have strong connections among themselves. The algorithm that was used is simple and uses DFS search as follows: It loops on the nodes and make DFS search on each node and mark the visited nodes from that node, then adding them to set as one component. If there is a node that is visited, no DFS search is made. The results were as follows, there was one dominant component acquiring 18817 nodes and other components that are shown in the figure with their frequencies.

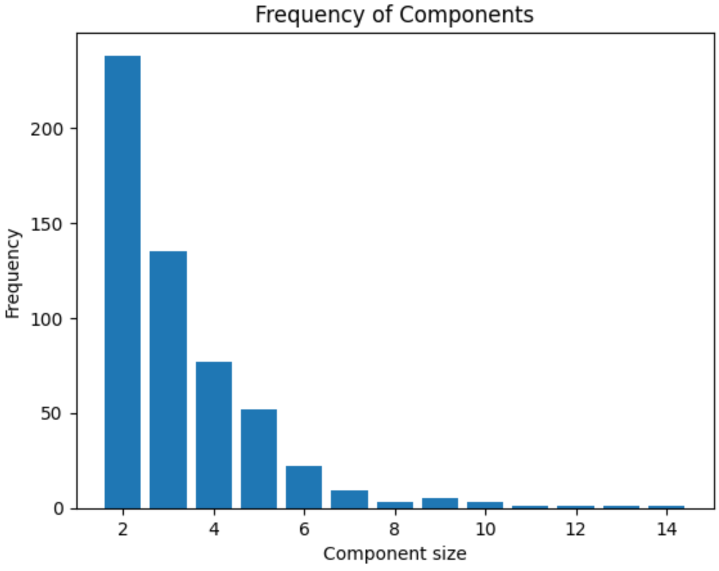

Figure 2: Components Frequencies

We see that there is a component that is dominant in the graph acquiring more than 90% of the nodes, which is a significant portion of the overall network. The other parts are concentrated in the range of 2-10 nodes. The giant component is introduced because of the people who worked in different companies, those are the nodes that connects the huge sub-graph with each other. The 2-in-size components are the most frequent components and they are for sure a person who have worked at only one company and there are no nodes in the graph worked in that same company.

Another interpretation is that if we have a dynamic data, the dominant components will increase its domination by acquiring and merging new or old components to it. And that will happen because new data will most probably has people who will work in some of those famous companies found in the dominant component.

3.5 Clustering coefficients

Clustering is another measure that describes the relationship between neighbours. A triplet is a terminology used to describe 3 nodes, if all the 3 nodes are connected (have 3 edges between each other) then it is known as closed triplet. And if they are not connected, they are said to be open triplets. Global Clustering Coefficient represents the ratio of the closed triplets with respect to the total triplets (open + closed ones). Another algorithm is for measuring the clustering coefficient for a specific node separately. It counts the number of edges between neighbours of that node and divide it by the maximum possible number of edges, which is $\binom{n}{2}$

For our graph the coefficient was 0.0. And that is because The nature of our graph is that there is no relation between the workers and each other or the companies and each other so there is no triplets in our graph. A clustering coefficient of 0.0 we got represents this fact that the graph has no closed triplets or triangles. In other words, it indicates that there are no connected triples of nodes that form a closed loop.

The clustering coefficient measures the degree to which vertices in a graph tend to cluster together. A value of 0.0 means that there are no local clusters within the graph. Each node's neighbors are not interconnected or forming triangles with each other.

It's important to note that a clustering coefficient of 0.0 does not necessarily mean that the graph is completely disconnected or lacks any structure. It simply indicates that there are no local clusters of nodes that exhibit high levels of connectivity among their neighbors.

3.6 Density analysis

Density analysis is similar to clustering but the density deal with all the graph. The density of the graph is the ratio between all the present edges in the graph and the maximum possible number of edges in the graph (The maximum number is n(n-1) where n is the number of nodes). Our graph's density is 0.000138 and that is extremely low but predictable value as the graph does not have information about the connections between the employees and each other or the companies and each other. The density of the dominant component alone was also calculated and it is 0.000159. It is more than the whole graph and that is obvious because this is the biggest past of the graph.

3.7 Community discovery

Community discovery analysis, also known as community detection or clustering analysis, is a process of identifying groups or communities within a network or data-set. It aims to uncover cohesive subgroups or clusters of nodes (vertices) that have a higher density of connections within the group compared to connections with nodes outside the group.

Community discovery analysis is commonly applied in various fields, including social network analysis as we do in our research, biology, computer science, and data mining. It helps in understanding the structural organization, functional relationships, and information flow within complex systems.

After, representing the data as a network, where nodes represent entities (employees , organizations, genes, etc.) and edges represent relationships or connections between them. We applied communities detection algorithms.

Various algorithms and techniques have been developed to detect communities within networks. These algorithms utilize different principles, such as optimizing modularity, maximizing edge betweenness, or minimizing graph cuts. Commonly used algorithms include the Girvan-Newman algorithm, Louvain method, Markov Clustering Algorithm (MCL), and hierarchical clustering.

In this research, we used Louvain method to detect communiies. Louvain method is a popular algorithm for community detection in networks. It is known for its efficiency and ability to detect communities at multiple scales. The algorithm is based on the concept of modularity, which measures the quality of a network partition into communities.

The algorithm starts by assigning each node in the network to its own separate community. This means that initially, each node forms its own community, Then attractively optimizes the modularity of the network by merging and rearranging communities. The goal is to find a partition of the network that maximizes the modularity value.

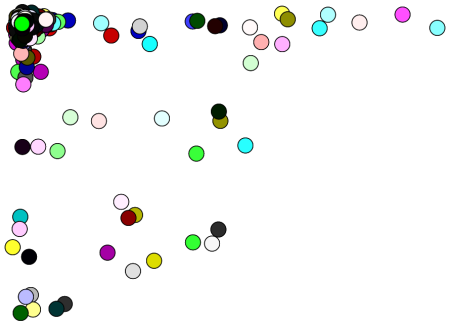

Figure 3: communities distribution

Applying Louvain method for our linkedIn data result in generating 617 community, communities size range from 2 to 1925. there is a large number of detected communities with size of 2 which means the network isn't strongly connected. Fig(3) shows relation between communities and how they are distributed around the network.

The larger communities (those with sizes greater than 500) represent the core or central groups within the network. These communities likely have a higher level of interconnectedness and influence compared to the smaller communities.

Largest community in the network contains around 1925 with 657 companies and 1268 employees this means that most peoples who changed their work companies work in one of companies in this community indicating that this set of companies related to the same field and have similar communities like (Standard Bank Group, Savvy Finance, Maestrano, and else in finance field)

4 Conclusion

Graph analysis is a valuable tool that provides meaningful insights into our data. By exploring the communities within the graph, we have uncovered valuable information that can be utilized to build recommendation systems for individuals or companies seeking new job opportunities. This allows us to suggest improved prospects for employment, matching people with suitable positions. Additionally, by considering the degree analysis and community structure, we can identify significant companies. Leveraging this information, we can recommend relevant companies to individuals within those communities or in closely connected communities. This targeted approach enhances the effectiveness of our recommendations, ensuring a better match between individuals and companies. Overall, graph analysis opens up new avenues for making informed recommendations in the job market, benefiting both job seekers and companies.

Acknowledgments

We would like to express our sincere gratitude to Dr. Walid Gomaa for his invaluable guidance, expertise, and support throughout the entire research process of this paper.

Also We extend our heartfelt appreciation to Eng. Ziad Shokry for his invaluable assistance in data collection and guidance in selecting the best analysis tools for our graph analysis research.