Social Network Analysis of NYC Jobs Post

Instructor: Prof. Feng Mai

Group 6: Bowen Lu, Zihan Chen, Jingmiao Shen, Yuan Zhou

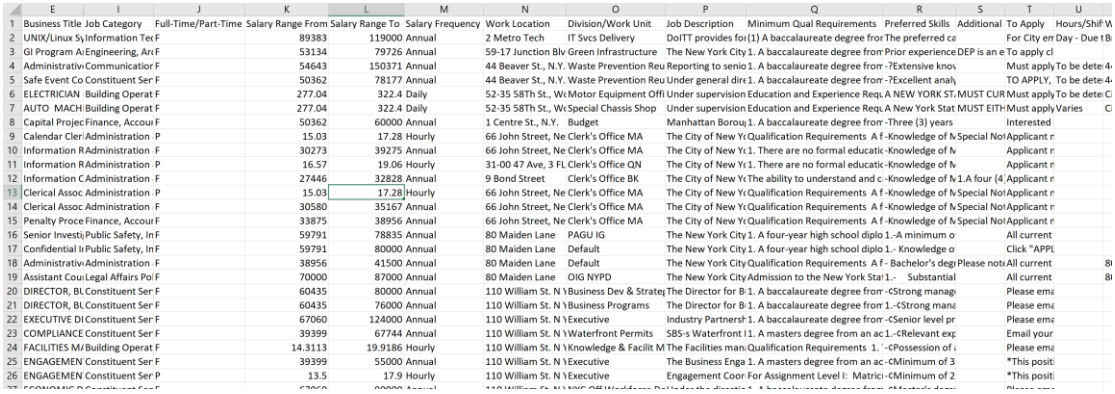
Background

As international students, we always want to gain more experiences from both the school and the industry. A lot of people are carrying the dream that landing a full-time job here in the US to choose to spend time and money studying far away from their own countries. The summer is coming which means there are many opportunities as well as much time for students to look for summer internship or a full-time job. Therefore, it is important to answer several popular questions with regards to job hunting: what kind of experience or skills are required in the job posting? What is the best way to achieve our career goals?

With that reality in mind, we plan to study the posted jobs in New York City for instance, analyzing the connections of jobs based on the required professional skills. As a result, we define the job positions as nodes and define the similarity of preferred skills of two positions as links for the network analysis.

Data Source

Our dataset comes from the NYC Open Data (https://opendata.cityofnewyork.us/) which is a popular data source supported by the Department of Citywide Administrative Services (DCAS). To be specific, the dataset contains 3,295 rows and 28 columns, including “Business Title”, “Agency”, “Job Category”, “Job description”, “Preferred Skills”, and “Work Location”, etc.



	E	I	J	K	L	M	N	O	P	Q	R	S	T	U
1	Business Title	Job Category	Full-Time/Part-Time	Salary Range From	Salary Range To	Salary Frequency	Work Location	Division/Work Unit	Job Description	Minimum Qual Requirements	Preferred Skills	Additional	To Apply	Hours/Shif
2	UNIX/Linux Sy	Information Tec	F	89383	119000	Annual	2 Metro Tech	IT Svcs Delivery	DoITT provides for	(1) A baccalaureate degree fror	The preferred ca		For City em	Day - Due t
3	GI Program A	Engineering, Arc	F	53134	79726	Annual	59-17 Junction Blv	Green Infrastructure	The New York City	1. A baccalaureate degree fror	Prior experience	DEP is an e	To apply cl	
4	Administrativ	Communicatior	F	54643	150371	Annual	44 Beaver St., N.Y.	Waste Prevention Reu	Reporting to senio	1. A baccalaureate degree fror	-?Extensive kno		Must apply	To be dete
5	Safe Event Co	Constituent Ser	F	50362	78177	Annual	44 Beaver St., N.Y.	Waste Prevention Reu	Under general dire	1. A baccalaureate degree fror	-?Excellent anal		TO APPLY,	To be dete
6	ELECTRICIAN	Building Operat	F	277.04	322.4	Daily	52-35 58Th St., W	Motor Equipment Offi	Under supervision	Education and Experience Requ	A NEW YORK ST	MUST CUR	Must apply	To be dete
7	AUTO MACH	Building Operat	F	277.04	322.4	Daily	52-35 58Th St., W	Special Chassis Shop	Under supervision	Education and Experience Requ	A New York Stat	MUST EITH	Must apply	Varies
8	Capital Projec	Finance, Accour	F	50362	60000	Annual	1 Centre St., N.Y.	Budget	Manhattan Borou	1. A baccalaureate degree fror	-Three (3) years		Interested	
9	Calendar Cler	Administration	P	15.03	17.28	Hourly	66 John Street, Ne	Clerk's Office MA	The City of New Yo	Qualification Requirements	A f -Knowledge of N	Special Not	Applicant r	
10	Information R	Administration	F	30273	39275	Annual	66 John Street, Ne	Clerk's Office MA	The City of New Yo	1. There are no formal educatic	-Knowledge of N		Applicant r	
11	Information R	Administration	P	16.57	19.06	Hourly	31-00 47 Ave, 3 FL	Clerk's Office QN	The City of New Yo	1. There are no formal educatic	-Knowledge of N		Applicant r	
12	Information C	Administration	F	27446	32828	Annual	9 Bond Street	Clerk's Office BK	The City of New Yo	The ability to understand and c	-Knowledge of N	1.A four (4	Applicant r	
13	Clerical Assoc	Administration	P	15.03	17.28	Hourly	66 John Street, Ne	Clerk's Office MA	The City of New Yo	Qualification Requirements	A f -Knowledge of N	Special Not	Applicant r	
14	Clerical Assoc	Administration	F	30580	35167	Annual	66 John Street, Ne	Clerk's Office MA	The City of New Yo	Qualification Requirements	A f -Knowledge of N	Special Not	Applicant r	
15	Penalty Proce	Finance, Accour	F	33875	38956	Annual	66 John Street, Ne	Clerk's Office MA	The City of New Yo	Qualification Requirements	A f -Knowledge of N	Special Not	Applicant r	
16	Senior Investi	Public Safety, In	F	59791	78835	Annual	80 Maiden Lane	PAGU IG	The New York City	1. A four-year high school diplo	1.-A minimum o		All current	
17	Confidential I	Public Safety, In	F	59791	80000	Annual	80 Maiden Lane	Default	The New York City	1. A four-year high school diplo	1.- Knowledge o		Click "APPL	
18	Administrativ	Administration	F	38956	41500	Annual	80 Maiden Lane	Default	The New York City	Qualification Requirements	A f - Bachelor's degr	Please not	All current	
19	Assistant Cou	Legal Affairs Po	F	70000	87000	Annual	80 Maiden Lane	OIG NYPD	The New York City	Admission to the New York Sta	1.- Substantial		All current	
20	DIRECTOR, BL	Constituent Ser	F	60435	80000	Annual	110 William St. N Y	Business Dev & Strate	The Director for B	1. A baccalaureate degree fror	-CStrong manag		Please ema	
21	DIRECTOR, BL	Constituent Ser	F	60435	76000	Annual	110 William St. N Y	Business Programs	The Director for B	1. A baccalaureate degree fror	1.-CStrong mana		Please ema	
22	EXECUTIVE DI	Constituent Ser	F	67060	124000	Annual	110 William St. N Y	Executive	Industry Partnersh	1. A baccalaureate degree fror	-CSenior level pr		Please ema	
23	COMPLIANCE	Constituent Ser	F	39399	67744	Annual	110 William St. N Y	Waterfront Permits	SBS-s Waterfront i	1. A masters degree from an ac	1.-CRelevant exp		Email your	
24	FACILITIES MA	Building Operat	F	14.3113	19.9186	Hourly	110 William St. N Y	Knowledge & Facilit M	The Facilities man	Qualification Requirements	1. -CPossession of		Please ema	
25	ENGAGEMEN	Constituent Ser	F	39399	55000	Annual	110 William St. N Y	Executive	The Business Enga	1. A masters degree from an ac	-CMinimum of 3		*This positi	
26	ENGAGEMEN	Constituent Ser	P	13.5	17.9	Hourly	110 William St. N Y	Executive	Engagement Coor	For Assignment Level I: Matric	-CMinimum of 2		*This positi	

Exploratory Data Analysis

To uncover insights from the dataset, it is better to conduct some EDA firstly.

At the beginning, a bar chart (Figure 1.1) is generated about the frequency of job titles. As we can see, the top three are “summer college intern”, “confidential investigator” and “public health inspector”, which gives a general view about the targeted job position.

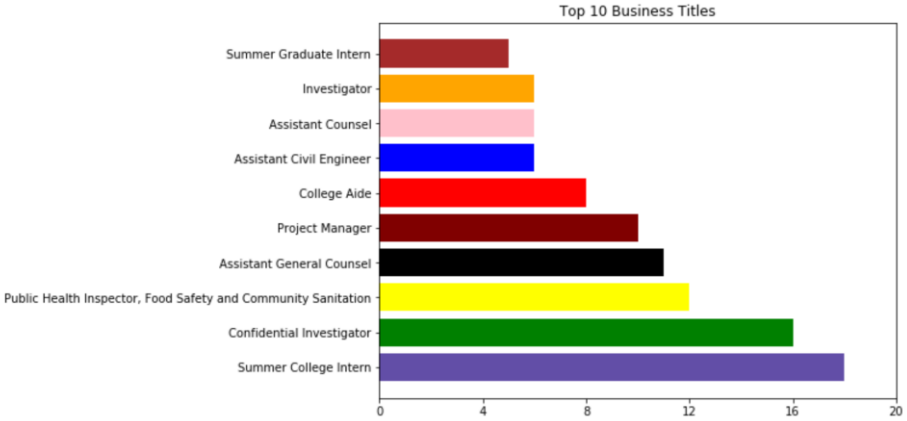


Figure 1.1

Figure 1.2 is a bar chart about the frequency of job categories. The top three are “Engineering”, “Technology” and “Health”. They can be the most concerned industries in further studies.

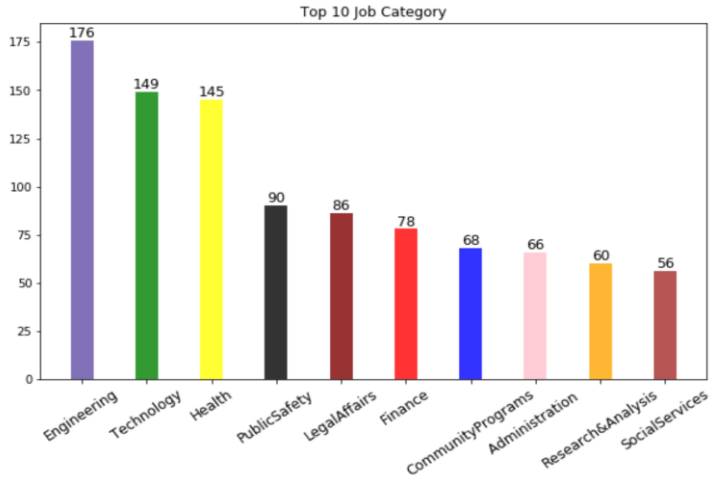


Figure 1.2

Figure 1.3 is a histogram showing the distribution of job salaries (annual). Here, we find the majority ranges from 50,000 USD to 100,000 USD.

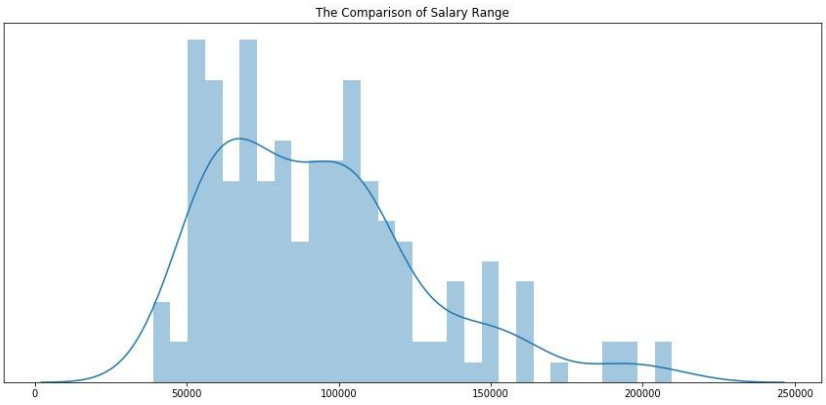


Figure 1.3

Figure 1.4 is a heat map showing the work location of jobs. It is clear that Manhattan is the most popular choice of the locations for these jobs.

Figure 1.4

Figure 1.5 is a radar plot summarizing the general requirement among the top three industries. We attribute the preferred skills of jobs of into five categories: “Domain knowledge”, “Tool”, “Experience”, “Education Background” and “Soft Skill”. All the three job categories have a large proportion that requires domain knowledge. The jobs belongs to technology have a less proportion that requires soft skills compared with other two categories.

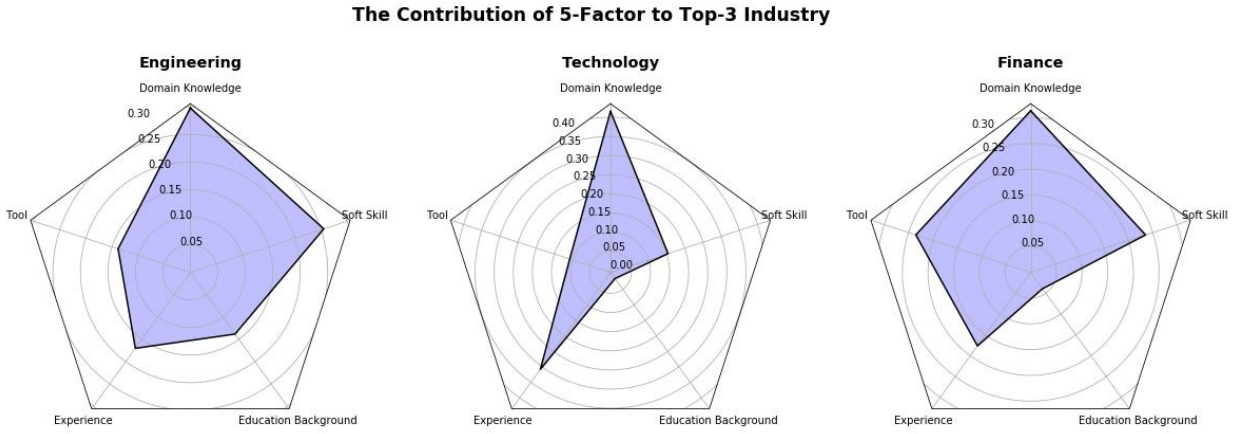


Figure 1.5

Network Analysis

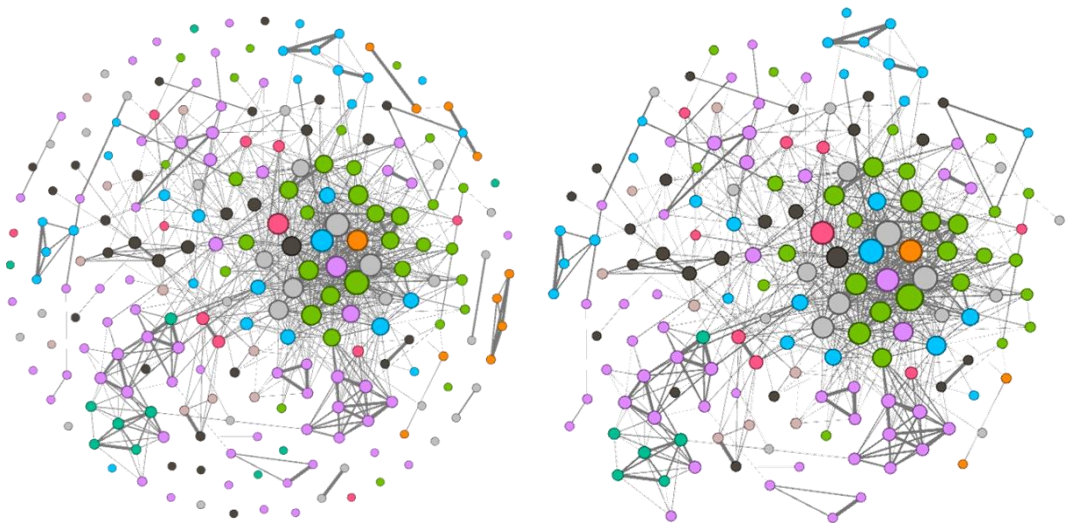
I. Overview

After dropping all the records which have missing values or messy codes in columns, 225 job positions are retained. As a result, we just utilized them as 225 nodes to build the whole

network for analysis. Their attributes are also defined including “Job title”, “Agency”, “Job field”, “Full/Part time”, “External/Internal”, etc.

The similarity of “Preferred skills” of two job positions is applied to define the link between two nodes. Several NLP methods are adopted to process the text of “Preferred skills” including removing numbers & punctuations & stop-words & whitespace, and stemming the terms. After generating the document-term matrix, the cosine similarity between each two documents of “Preferred skills” can be obtained. A threshold is set that the link can be reserved only if the cosine similarity larger than 0.4 (the value is defined as the weight of link). Otherwise, the link will be deleted.

As a result, the complete network is generated (Figure 2.1-a), being composed of 225 nodes and 927 weighted but undirected links. By eliminating the isolated nodes, the giant component of this network consists of 169 nodes and 913 links (Figure 2.1-b). The further network analysis is mainly based on the giant component.



a **b**

Figure 2.1

1. Network-wise analysis

Figure 2.2 is the giant component of the whole network, with the job title as the label of nodes. The color is distinguished by different job categories that job positions belong to. The size of nodes is determined by the degree of nodes. The degree centralization of this network is 0.0747, showing that it is a low centralization network with more even distribution of similar jobs.

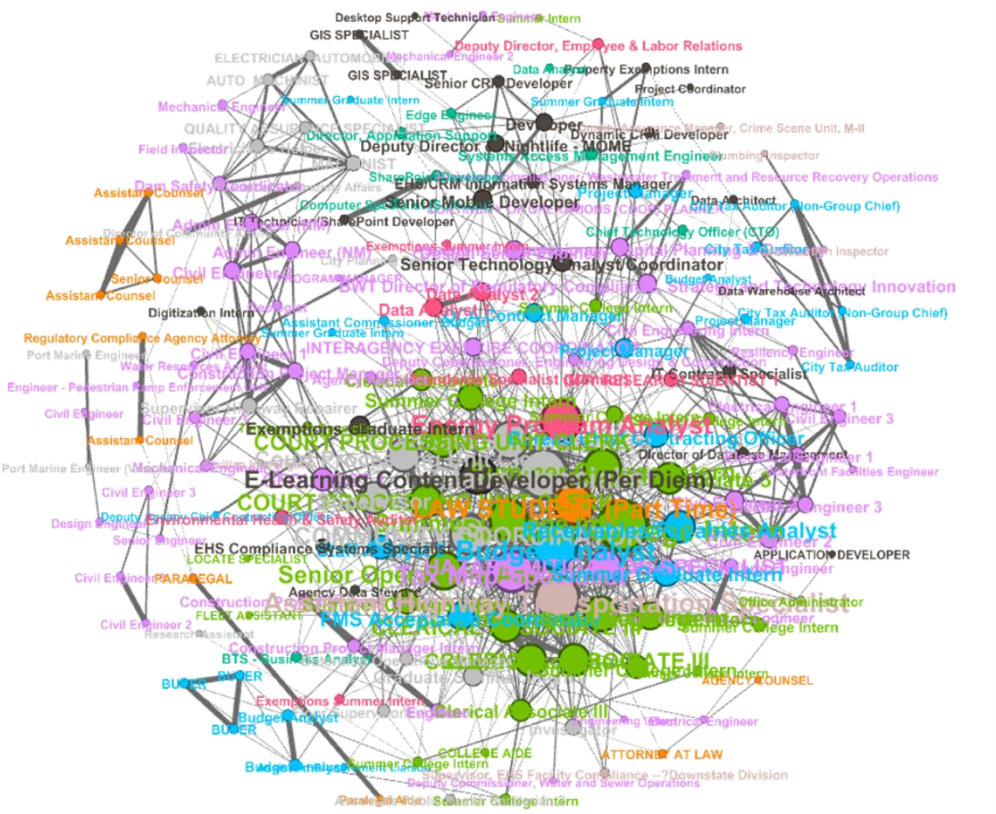


Figure 2.2

Due to the threshold that only the links which corresponding cosine similarities larger than 0.4 can be reserved, theoretical existence of a large group of links is eliminated. That is the reason why the density of this network is only 0.0545. However, the idea cannot be come up with that this network is weakly connected. The transitivity (clustering coefficient) of the network is 0.4955. Therefore, it is quite common in this network that if the preferred skills of two jobs are both similar to that of the third job, the first two jobs are likely to be similar in the preferred skills.

As for the homophily of this network, we calculate the assortativity of node degrees and the assortativity of some attributes of nodes. The assortativity of node degrees is 0.2939, reflecting a not significant bias of connections between nodes with similar degrees. The homophily about the job's posting type (external/internal) or status (full/part time) of the network is much more not significant, since the assortativities about these two attributes of nodes are 0.0675 and -0.0153.

2. Node-wise analysis

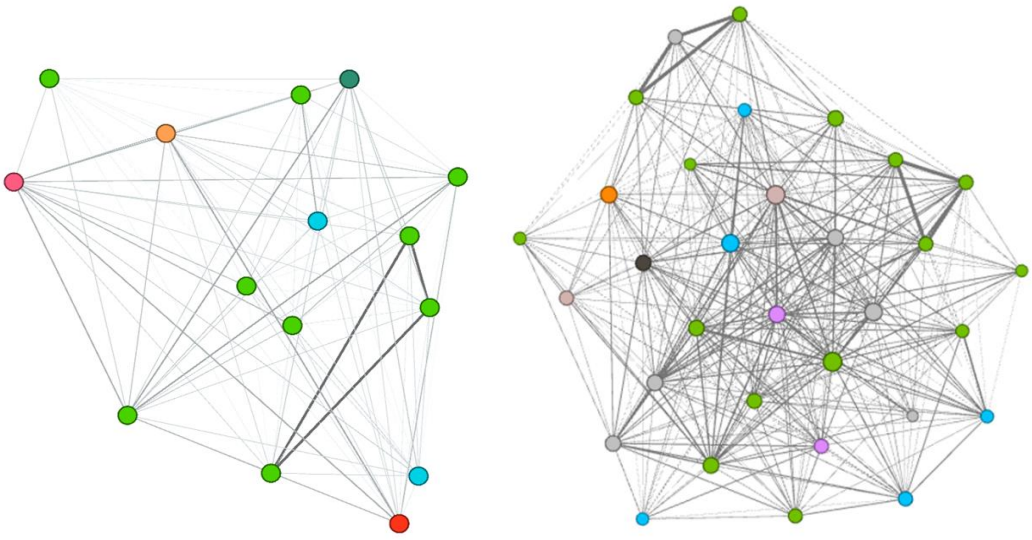
One of the essential node-wise analysis is the centrality. We apply four kinds of centrality measurement to calculate and compare the normalized and non-normalized centralities of nodes. The nodes with highest centrality are displayed in Table 2.1.

Summer college intern has the highest degree centrality and highest eigenvector centrality, since this position is the starting position in many job categories. Without strict requirement of professional skills, it performs as a hub node linking with many important jobs. However, claim specialist has the highest closeness centrality and betweenness centrality, which performs as a bridge node building a connection between jobs in different categories.

Node	Normalize=F	Normalize=T
Degree		
Summer college intern	48	0.2609
Betweenness centrality		
Claim Specialist	1531	0.0909
Closeness centrality		
Claim Specialist	3.194×10^{-4}	0.0588
Eigenvector centrality		
Summer college intern	1	

Table 2.1

We also study the nodes having complete mutuality with each other. As taught in class, a clique is a group of nodes that all adjacent to one another. In our network, the largest clique has 15 nodes (shown in Figure 2.3-a). Each node has same degree = 14, and the amount of links is (15*14)/2 = 105. If screening all the nodes whose degree not smaller than 14, a K-core (K=14) is formed (shown in Figure 2.3-b). The K-core has 32 nodes and 352 links.



a **b**

Figure 2.3

II. Partial View

In this part, we conduct partial studies of the network, combining the knowledge of network analysis and the real-life situation together so that job seekers can find something helpful with the assistance of the network we build.

1. Single job category

This is for those who prefers changing jobs within the same category. Figure 3.1 is the subnetwork composed by the nodes belonging to the job category of Engineering.

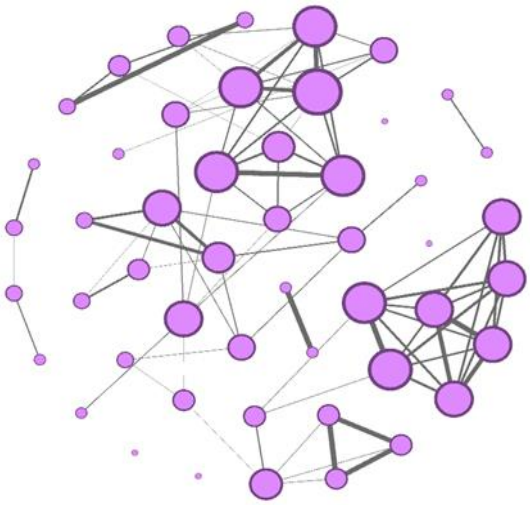


Figure 3.1

There are two features exemplifying the effective usage of this subnetwork.

1) Hub:

   Hub is the node with high degree. In other words, hub is a node that connects with plenty of other nodes that greatly exceeds the average. Intuitively, we think the hub nodes refers to the jobs which has a bigger chance to be basic jobs than the jobs with strict requirement of skills. The distribution of jobs of different professional levels is just like a pyramid: the base level is filled with jobs without strict requirement of skills, but they are the entry position of various industries; the top level is composed of jobs with the highest level of specialization, difficult to switch to another industry.

2) Bridge:

   Bridge is the node that if it was deleted, then the end points will belong to different component. In the network we study, the bridge nodes refer to transitional jobs. For example, Job A requires knowledge on computer science and Job B requires knowledge on architecture. These two nodes have no direct link but they are both adjacent to Job C. If a person in A wants to switch to B, he/she can acquire preferred skills by doing C first, which will boost the chance to be admitted by B.

In order to explain in detail about how the network can guide the real-life job searching, we select a part of nodes from Figure 3.1 and retain the links between those nodes (shown in Figure 3.2).

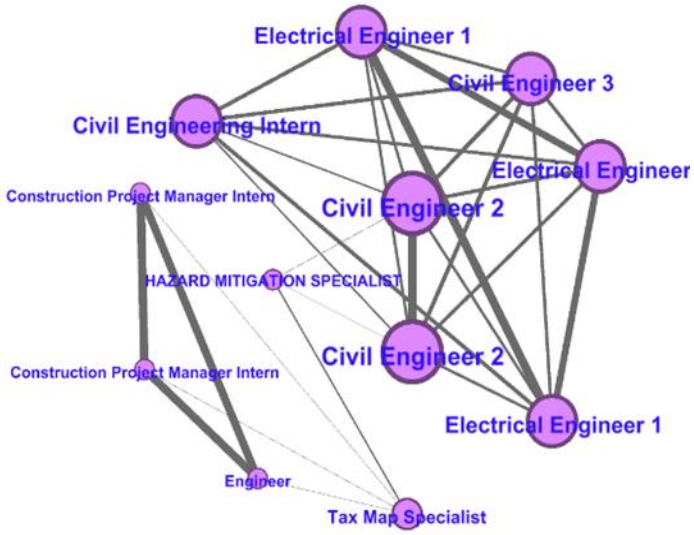


Figure 3.2

The nodes with the title of civil/electrical engineer form a cluster in the upper right part of this small network. While the nodes related to management (including the node "Engineer" because it is hired by the department of construction management) form a cluster in the left lower part of this small network. The nodes of one cluster have no directly link with the nodes of another cluster. But two clusters are connected by the node "Tax Map Specialist" and "Hazard Mitigation Specialist". Supposing that a civil/electrical engineer plans to switch to the managing job position, it is not likely to successfully switch to the construction project manager intern directly since there are little similarity between them. However, the engineer can firstly apply for the Hazard Mitigation Specialist and Tax Map Specialist to accumulate the knowledge and ability in finance and risk management. Then the probability of switching to the project manager will arise significantly.

2. Multiple job categories

This is for those who prefers changing jobs to different companies or different industries. Figure 3.3 is the subnetwork composed by the nodes belonging to three job categories: ① Administration & Human Resource (green nodes); ② Finance, Accounting & Procurement (blue nodes); ③ Technology, Data & Innovation (black nodes).

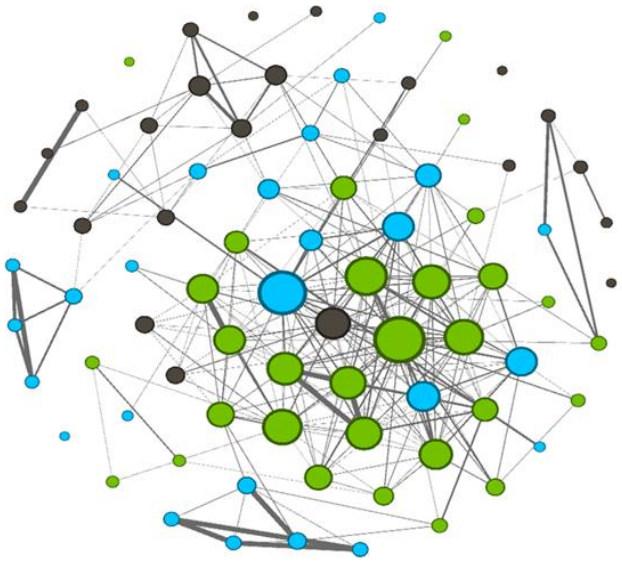


Figure 3.3

Although the nodes are divided into three categories, the links are determined by the similarity of preferred skills, not significantly affected by the difference of job category. Therefore, the nodes belonging to finance can directly link with the nodes belonging to technology as long as the skills requirement is similar. The features in the network of single job category can also be analyzed in the network of multiple job categories without variations to a large extent. So, they will not be repeated.

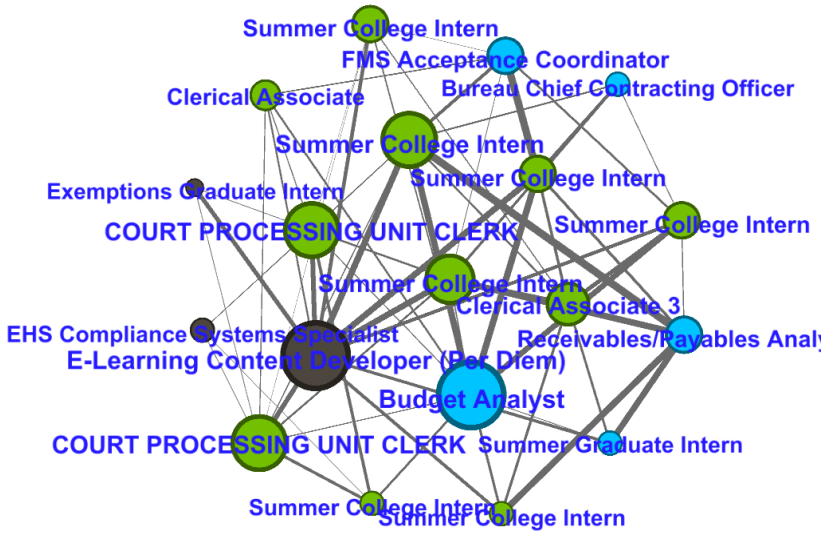


Figure 3.4

Figure 3.4 composes of a part of nodes of certain companies from Figure 3.3 and retain the links between those nodes. Supposing that a person working as Exemptions Graduate Intern (a small black node located at the upper left in Figure 3.4) want to become a Budget Analyst of another company, he/she need to cross the job categories. The node of Exemptions Graduate Intern has low degree but a high closeness centrality since it is

adjacent to the high-degree nodes E-learning Content Developer and Court Processing Unit Clerk. Thus, these important neighbors are the first step for the Exemptions Graduate Intern to become a Budget Analyst. Although the Exemptions Graduate Intern has strong connection with the E-learning Content Developer since they are in same category and having high similarity for preferred skills, the career path from Exemptions Graduate Intern to Budget Analyst via E-learning Content Developer is not the shortest because E-learning Content Developer is not adjacent to Budget Analyst (otherwise, there must be a strong tie between Exemptions Graduate Intern and Budget Analyst). The shortest path is via Court Processing Unit Clerk because it is both adjacent to Exemptions Graduate Intern and Budget Analyst (the length is 0.826).

III. Community Detection

For the whole network we generate in Figure 2.1-a, it's easy to observe that the nodes inside each subgroup of the large network usually have more tight connection between each other. We call this kind of subgroup "community", which means the nodes in the community always has links to other nodes more densely than to the nodes outside the community. We choose two algorithms to detect the communities: edge-betweenness algorithm and fast-greedy algorithm.

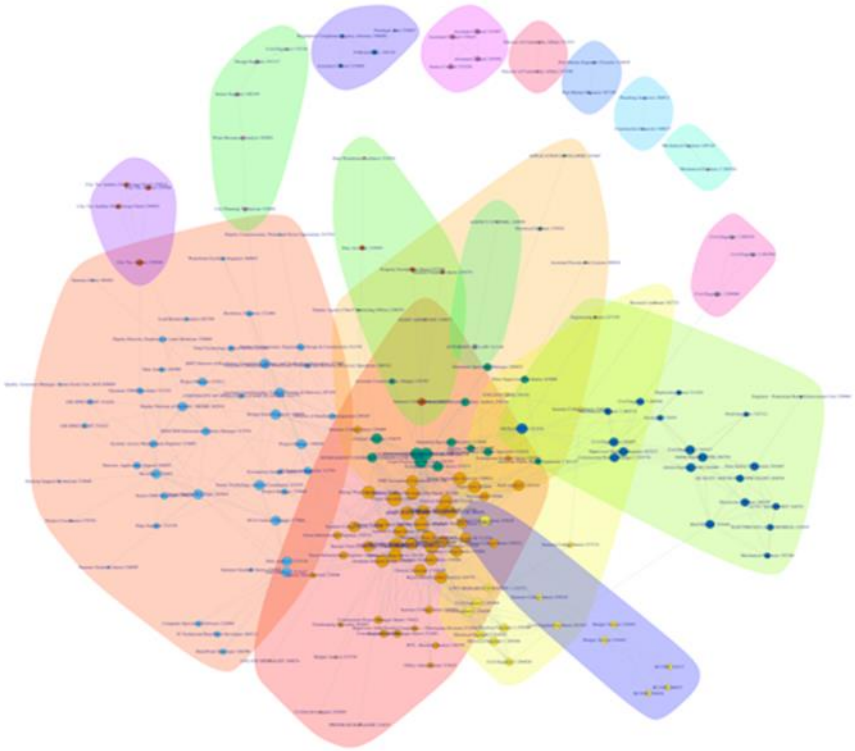


Figure 4.1

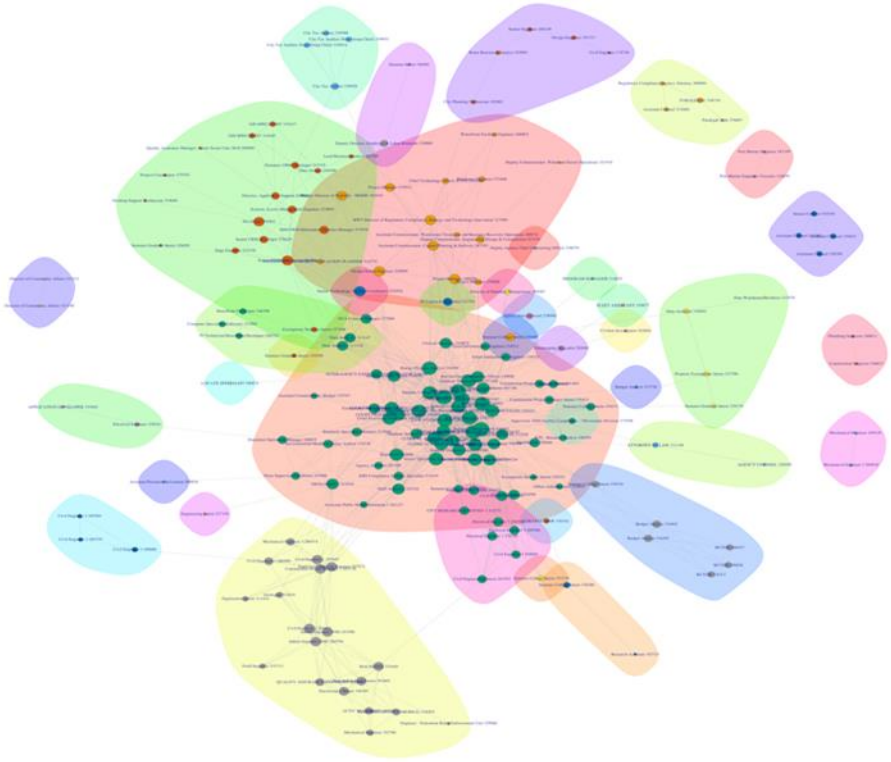


Figure 4.2

Figure 4.1 the community graph by using the edge-betweenness algorithm, while Figure 4.2 is the community graph by using the fast-greedy algorithm. The first observation we could find is that most of the big communities are not disjoint but overlapping, in both community detection methods. It means that there are some position performing as bridges, which connects different communities. A good example is the Summer College Intern. It not only connects with jobs like Civil Engineering, Computer Specialist, but also leads the way into Public Health and Finance. Therefore, the first conclusion we could get though community detection is that if you want to jump into a brand-new field, you should never miss the Intern opportunity in the summer.

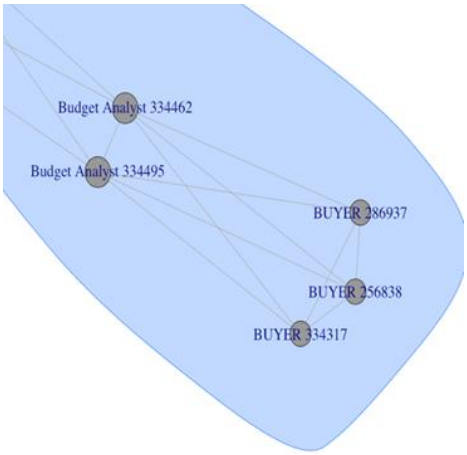


Figure 4.3

Then we tried to zoom in this picture. Figure 4.3 is a local view of one community in the Figure 4.1. This blue area consists of Budget Analyst and Buyer, indicating that a buyer always needs good budget skills. But in practical situation, it's not very common, because budget skills could be applied into far more ways. When turning into Figure 4.2, we found that Budget Analyst is not constricted in this small community anymore, but act as a key node into various communities with nodes like Civil Investigation, Office Administrator, even the Timekeeping Specialist.

Another interesting finding is the role of Data Analyst and Business Analyst. These two positions are popular nowadays, but many job seekers feel confused about the difference between them. Figure 4.4 is a part of Figure 4.1 containing Data Analyst, while Figure 4.5 is a part of Figure 4.1 containing Business Analyst.

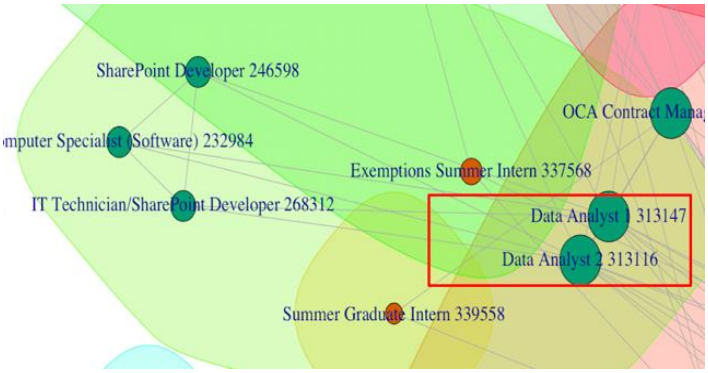


Figure 4.4

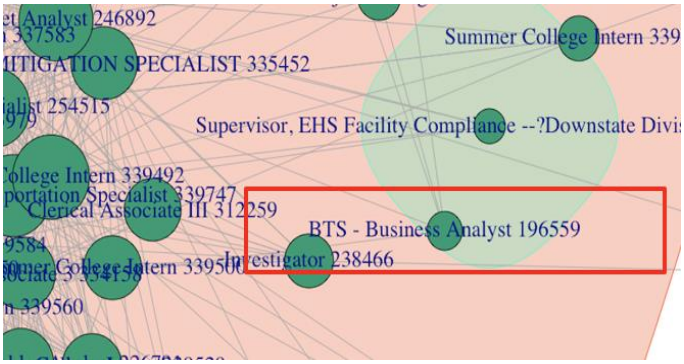


Figure 4.5

For data analyst, most of its neighbors in the green community are computer related, like software engineer and IT developer, which means good computer skills are required before applying the data analyst. For the business analyst, it locates in the biggest community in the network, which requires professional skills with low threshold, but wide armory. Therefore, a business analyst should handle the information from all kinds of jobs and situation, requiring higher level analysis ability. However, in Figure 4.2 generated by the fast-greedy algorithm, data analyst also link with nodes of various communities instead of computer science only, like budget analyst does.

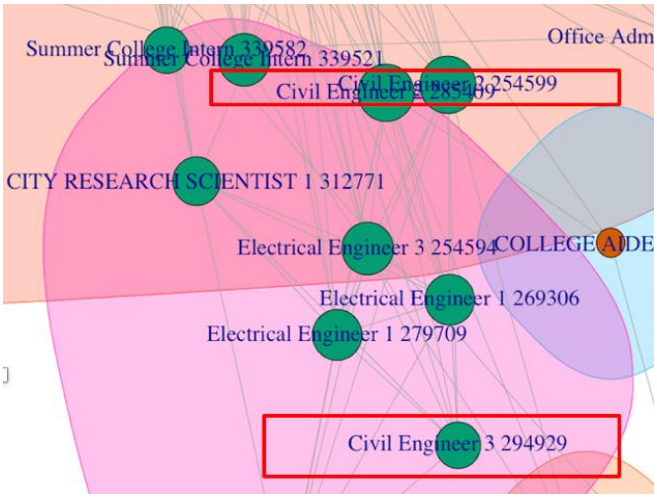


Figure 4.6

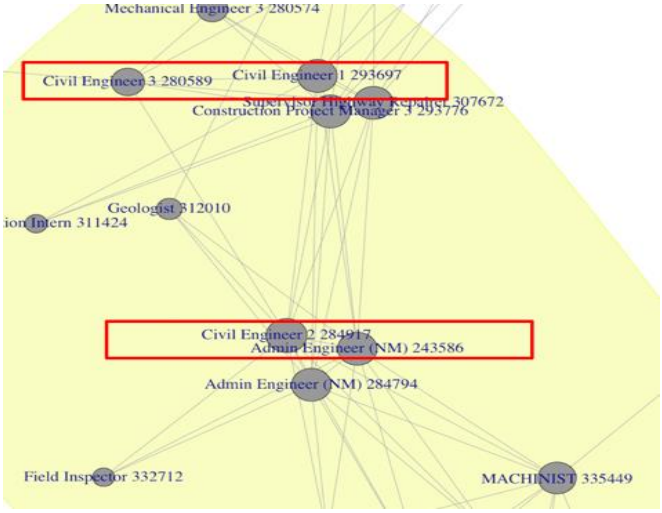


Figure 4.7

The next attractive discovery is a very powerful position, Civil Engineer. As shown in Figure 4.6 and 4.7 as parts of Figure 4.1, the civil engineer acts as an omnipotent key linking with nodes including Electrical Engineer, Construction Engineer or City Research Scientist. If you want to take part in the construction project, Civil Engineer is a good stepping stone. Also for the company, it's a wise choice to hire more Civil Engineer student, then assign them into different position to meet all kinds of demand.

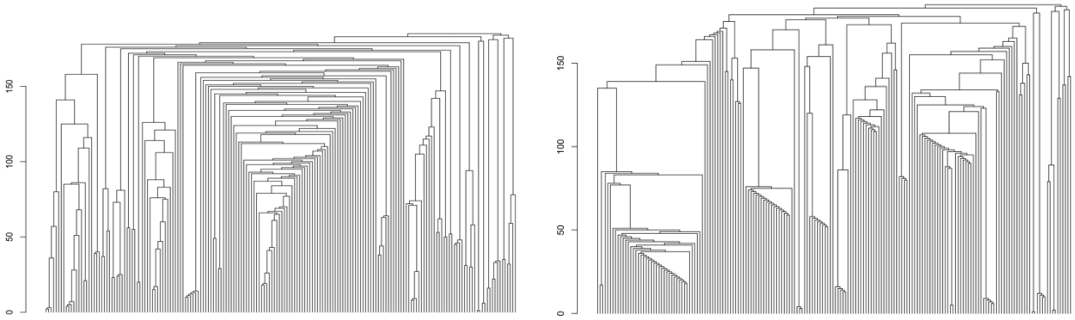


Figure 4.8

Finally, besides discovering the difference of communities received by two algorithms, we also think about the reason of such difference. Figure 4.8 shows the dendograms of the communities generated by the two algorithms (left one is the edge-betweenness, right one is the fast greedy). The structures of two dendograms are very distinct. As a result, when setting different thresholds towards two dendograms, we will get different results.

Business Implication

I. For job seekers:

Our networks build the links between job positions as nodes based on the similarity of text in the preferred skills of jobs. If there is a link between two nodes, it is possible to switch from one to another, and the larger weight of link the higher probability of success. Therefore, this network can provide job seekers with an effective path of career development in a visualized way. Following the method our project, job seekers can build a customized network consist of the jobs or job categories they may concern with specific scales. They can utilize the knowledge of network analysis to find the best path to the dream jobs and prepare for professional skills by participating the intermediate jobs. It is much more efficient and accurate than reading job descriptions and manually comparing them.

II. For companies

Our networks display the similarity relations among job positions, which is also worth making deep analysis for companies, especially the human resource department of companies. In order to hire people which are suitable for the skill requirement of their jobs, companies should focus on the preceding nodes which closely connecting with their job nodes and attract the employees of those preceding nodes to work for them. Meanwhile, companies should also pay attention to the competing companies' job nodes which are closely connected with or just adjacent to their job nodes, for fear that their employees are attracted by those competing companies.

Future Work

1. The links between nodes could be determined by a comprehensive and more accurate method. In this project, the links are defined by the cosine similarity of the text in preferred skills of job positions. However, two jobs with similar skill requirement can be disconnected due to other requirement like age, gender, status, working experience, residence place, etc. Considering that, the links and their weights can be determined by a new variable which

combining various factors together. It may be a score calculated by a linear equation of several variables, or it can be a success rate of switching from one job to another according to the historical records.

2. The network could be a directed network showing the progress of career development. In this project, when there is a link between a summer internship and a full-time job, both directions are allowed, but this will not happen in any cases. Everyone hopes to switch to a job with better conditions, higher salary, and brighter future when they decide to leave current jobs. Therefore, the network should not only display that which jobs are likely to switch to due to the required skills, but also give advices that which jobs can enhance their career. The direction can be determined by the salaries, other benefits, working hours, and company reputations, etc.