Social Network Analysis of Positions on LinkedIn

Shanliang Yao
Department of Computer Science and
Software Engineering
Xi'an Jiaotong-Liverpool University
Suzhou, China
Shanliang.Yao19@student.xjtlu.edu.cn

***Abstract*—LinkedIn is a business customer-oriented social network site especially known by its function of posting vacant positions. The idea proposed in this research is to extract information from the positions available in LinkedIn and identify the most valuable positions by centrality, density, connectedness, etc. Community detection and graph cluster are also used in this research to analyze the characteristics of the communities and relationships among positions. The high-frequency skills in the position have also been analyzed, which can help job seekers find the skills they need to better improve themselves in their careers.**

Keywords—social network analysis, LinkedIn, position, characteristics, skills

I. Introduction

A. Background

The increasing popularity of Online Social Networks (OSNs) is witnessed by a huge number of users acquired in a short time. Some social networking services have now attracted hundreds of millions of users, such as Facebook, Twitter, LinkedIn, etc. Users can share recent moments, talk with friends, learn online knowledge, and post position recruitments on social networks.

LinkedIn is a professional global workplace social platform and a social network for business customers. Company leaders or human resources publish their recruitment information on LinkedIn to attract professional talent on LinkedIn to apply. Recruitment information on LinkedIn is very representative and can reflect the overall market demand. We can learn the development of the positions in different aspects by analyzing the relationship among the positions in different industries, companies, and regions, using the technology in social network analysis to visually display the social graph. Skills needed for the position can also be mined to help the job seeker understand their career direction and purposefully improve their skills.

B. Problem Statement

a) Research problems

Generally, social network analysis is commonly used to analyze the relationship among people. Nodes are people, and edges are relationships among people. But how to analyze the positions in companies is worth thinking about. How to obtain representative position information also needs to be considered. Moreover, discovering the core skills of the position that need to be mastered is very important for job seekers.

b) Research questions

As the analysis of positions using social network techniques has some problems, this research aims to solve these problems. My research questions are addressed:

- RQ I: Among all given positions, what are the core and valuable ones?
- RQ II: How does social network analysis function in detecting the communities in positions and analyzing the differences between the communities?
- RQ III: Does social network analysis help in finding the most needed skills in positions under the same title?

II. Literature Review and Proposed Methods

This chapter first reviews the progress of position analysis and LinkedIn analysis. Then, it researches different technologies in the field of social network analysis. After that, it studies some algorithms on community detection and graph cluster analysis. And finally, by summarizing all, it identifies and pins down some key issues and proposed methods that need to be addressed in my research.

A. Progress of Position Analysis and LinkedIn Analysis.

The position analysis pares the responsibilities of a position down to the core functions necessary to successfully perform the work. The position analysis is useful in providing an overview of the fundamental requirements of any position.

In Zhang's study, the coordinate transformation matrix is used to represent the position of the moving platform [12]. And it presents the closed-form solution of the forward position analysis. A case study of workplace use of Facebook and LinkedIn points out that LinkedIn is developing faster and faster, which brings great convenience to people 's social networking and job sharing [13].

Another paper is to identify the elements of a LinkedIn profile that hiring professionals focus on most, and then examine LinkedIn profiles in terms of these identified elements across different industries [14]. This provides some ideas for me to analyze the structure of the position page.

B. Techniques of Social Network Analysis

Centrality is a concept commonly used in social network analysis to express the degree to which a point in a social network is at the center of the entire network. Density can be used to describe the density of interconnected edges between nodes. It is defined as the ratio of the actual number of edges in the network to the upper limit of the number of edges that can be accommodated.

1) Degree centrality is defined as the degree of the node that is the number of edges connecting the node. The greater the degree of centrality, the stronger the ability of the node to communicate directly with other nodes.

2) Closeness centrality measures the centrality of nodes in the network based on the distance between nodes. The smaller the average shortest path between a node and all other nodes in the network, the greater the centrality of the node.

3) Betweenness Centrality is defined as the proportion of the number of paths passing through the node among all the shortest paths in the network. The centrality of intermediary degree reflects the dependence degree of other nodes to reach other nodes of the network through the node and reflects the different nodes in the social network.

4) Density measures the degree of interconnection between N objects in the overall network, and its value ranges from 0 to 1. Generally speaking, the greater the density of the overall network, the greater the influence that the network may have on the attitudes and behaviors of the actors.

C. Algorithms of Social Network Analysis

1) Community structure by multi-level optimization of modularity is a method to extract communities from large networks [8]. The method is cluster_luvain [2] which means that in the Louvain method of community detection, first small communities are found by optimizing modularity locally on all nodes, then each small community is grouped into one node and the first step is repeated. The method is similar to the earlier method by Clauset, Newman and Moore [9] that connects communities whose amalgamation produces the largest increase in modularity.

2) Community detection based on edge betweenness is another method for community detection. The edge betweenness score of an edge measures the number of shortest paths through it [16]. The function of this method performs this algorithm by calculating the edge betweenness of the graph, removing the edge with the highest edge betweenness score, then recalculating edge betweenness of the edges and again removing the one with the highest score, etc.

3) Community detection based on spread labels means that the community attribution of each point is determined by the label of the adjacent node, and the label that appears most among the adjacent nodes is the label of that node [17]. In the beginning, each node has a unique label, and then the densely linked node labels are grouped, and then iterate continuously until these nodes with the same label form a community.

4) The kernel function can be regarded as a nonlinear transformation, which increases the separability of the input data by mapping the input data into a new high-dimensional space. Kernel k-means algorithm applies the same trick as k-means but with one difference that here in the calculation of distance, kernel method is used instead of the Euclidean distance [19].

5) The Highly Connected Subgraphs (HCS) Clustering algorithm is an algorithm based on graph connectivity for cluster analysis [10]. It works by representing the similarity data in a similarity graph and then finding all the highly connected subgraphs. In a similarity graph, for a given number of vertices, the more edges there are, the higher the similarity between such a set of vertices. In other words, if we try to disconnect the similarity graph by deleting edges, the more edges that need to be deleted before the graph is disconnected, the more similar the vertices of the graph.

D. Visualization of Social Network

Social network visualization can be constructed and implemented using the igraph package in R. Of course, there are many visualization tools for social network analysis without coding, such as Gephi and SocNetV.

Gephi is an open-source and free cross-platform JVM-based complex network analysis software [18]. It is mainly used in various networks and is favored because it is simple, easy to learn, and beautiful.

Social Network Visualizer (SocNetV) is a social network analysis and visualization application. we can draw a social network (graph/digraph) or load an existing one, compute cohesion, centrality, community and structural equivalence metrics and apply various layout algorithms based on actor centrality or prestige scores (i.e. Eigenvector, Betweenness) or dynamic models [11].

To build a tag cloud, the package named wordcloud2 in the R language is used. By passing words and the number of words, we can get the graph of the words.

E. Summary

While there has been much research on people using social network analysis, few researchers have taken positions into consideration. In this research, I applied social network analysis to position analysis to explore the relationship among positions, and to explore the most valuable positions.

Some proposed methods are used for analyzing the dataset, like degree centrality, closeness centrality, betweenness centrality and density methods. In addition, some algorithms are also used in this research, for example, Community structure by multi-level optimization of modularity, community detection based on edge betweenness, Community detection based on spread labels and HCS Clustering algorithm.

Moreover, some visualization packages and visualization software will also be used. In this way, the data analysis can be vividly displayed to help us understand the data.

III. Implementation and Application Demonstration

A. Research Design

a) Research Architecture

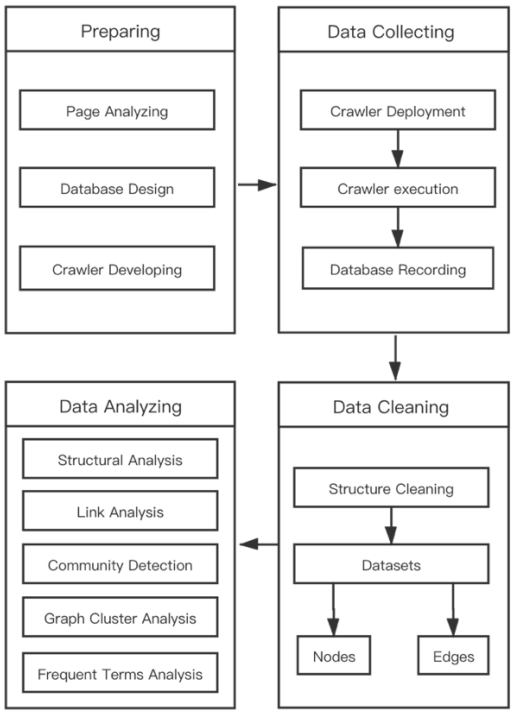


Fig. 1. Research architecture

As is described in Fig. 1, the implementation of the research includes preparing, data collecting, data cleaning and data analyzing.

I analyzed the structure of the source code of the position page on LinkedIn and found that its position list was obtained through ajax asynchronous request. The parameters of this API include job location, keywords, sorting method, etc. So, I can use the crawler to request different positions by passing different keywords.

b) Database Design

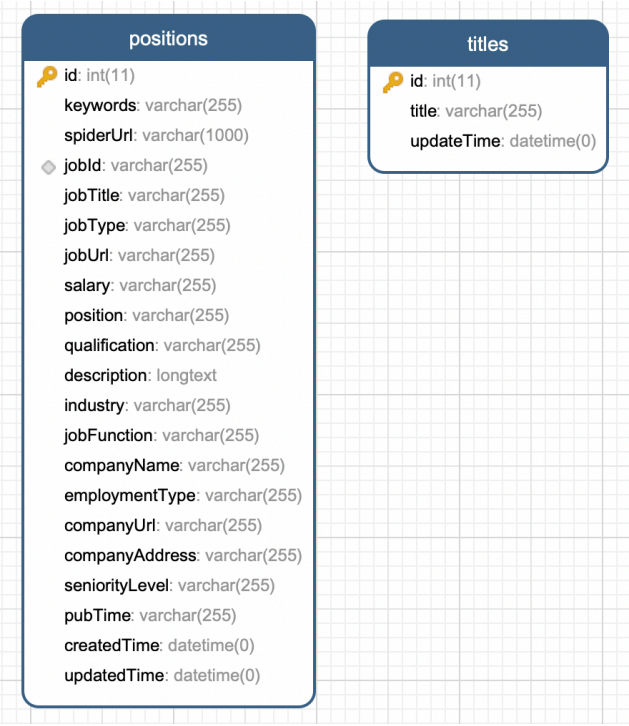


Fig. 2. Database for storing positions and titles

The database chosen in this research is MySQL with InnoDB storage engine. As is shown in Fig. 2, it contains table named positions and table named titles. For table positions, it has some attributed fields, like keywords, spider URL, job ID, job title, job type, description, industry, job function, etc. The table named titles is used to store titles of positions that will be crawled.

B. Crawler Developing

The crawler code is developed in python language and has been uploaded to my GitHub at https://github.com/yaoshanliang/linkedinSpider. It can be accessed by other developers who are interested in crawling. The crawler can automatically read the position title from the table title and then crawl the link for positions without manually modifying the code.

C. Experiment and Data Collecting

In this research, the position title named Social Network Analysis is used for collecting data. The crawler service is deployed on my server and can be automatically crawled each hour. The number of positions with keyword social network analysis currently obtained is more than 4,000 now. The dataset has also been published on my GitHub at https://github.com/yaoshanliang/linkedinAnalysis for other data analysts to do some research.

D. Data Cleaning

To analyze the connection among positions, edges need to be constructed. In the job function field, each position has one or more functions. Positions with the same function are connected by edges after separated. After that, the edges dataset is formed to be analyzed.

The following table shows the number of nodes and edges. Some nodes and edges are shown in Appendix 1 and Appendix 2.

TABLE I. NUMBER OF NODES AND EDGES

Graph	Total number	
	Nodes	*Edges*
position	100	1413

The above TABLE I shows the number of nodes and edges which will be used in the later analysis.

E. Data Analyzing

a) Structural Analysis

To vividly show the connection among positions, the plot function in R is used to draw the network. Firstly, a package named igraph should be included to show graphs. Then, the code reads the data in the positions and relations CSV files to build nodes and edges. Finally, the nodes and edges are passed into the function and specified as an undirected graph. The source code is in Fig. 3 below and the network visualization is shown in Fig. 4. Appendix 3 gives a clearer graph of the visualization.

```
library(igraph)

# Read files
positions <- read.csv("data/sna_positions.csv", header = T)
relations <- read.csv("data/sna_edges.csv", header = T)

# Construct the network
nodes <- data.frame(positions[, 1])
edges <- data.frame(from = relations[, 1], to = relations[, 2])
net <- graph_from_data_frame(edges, vertices = nodes, directed = FALSE)
plot(net, vertex.label.cex = 0.5, vertex.size = 10,
    main="Network visualization of the positions")
```

Fig. 3. Code to plot the network

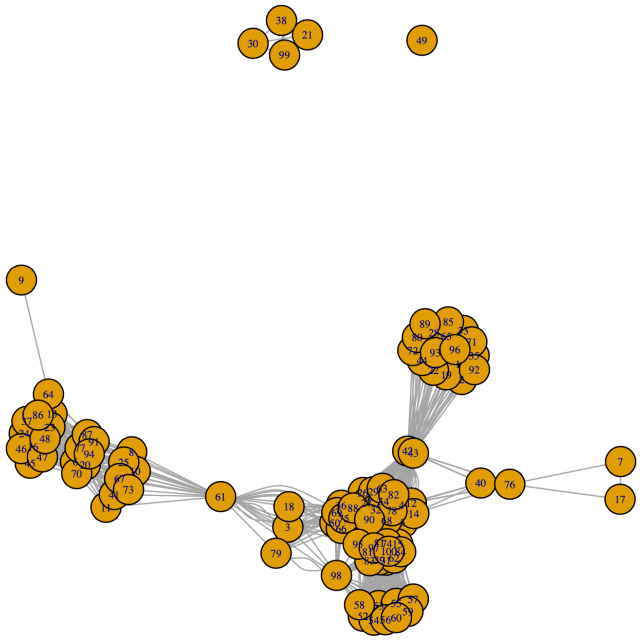


Fig. 4. Network visualization of the positions

Different measures of centrality for the assessment of positions are used, both at an individual and structural level. To analyze centrality, degree centrality is used, which is defined as the number of links incident upon a node. Betweenness Centrality is defined as the ratio of the number of paths passing through this node to the total number of shortest paths among all the shortest paths in the network. The centrality of betweenness reflects the dependence of other nodes to reach other nodes in the network through this node and reflects the different roles and positions of different nodes in social networks.

The degree centrality, closeness centrality and betweenness centrality are coded by R in Visual Studio Code. The code is as follows in Fig. 5 and the graph of the top 10 positions in the number of degrees are shown in Fig. 6.

```
library(igraph)
positions <- read.csv("data/sna_positions.csv", header = T)
relations <- read.csv("data/sna_edges.csv", header = T)

# Construct the network
nodes <- data.frame(positions[, 1])
edges <- data.frame(from = relations[, 1], to = relations[, 2])
net <- graph_from_data_frame(edges, vertices = nodes, directed = FALSE)

# Degree, closeness, betweenness
degree <- degree(net)
closeness <- closeness(net)
betweenness <- betweenness(net)
```

Fig. 5. Code to calculate degree, closeness and betweenness

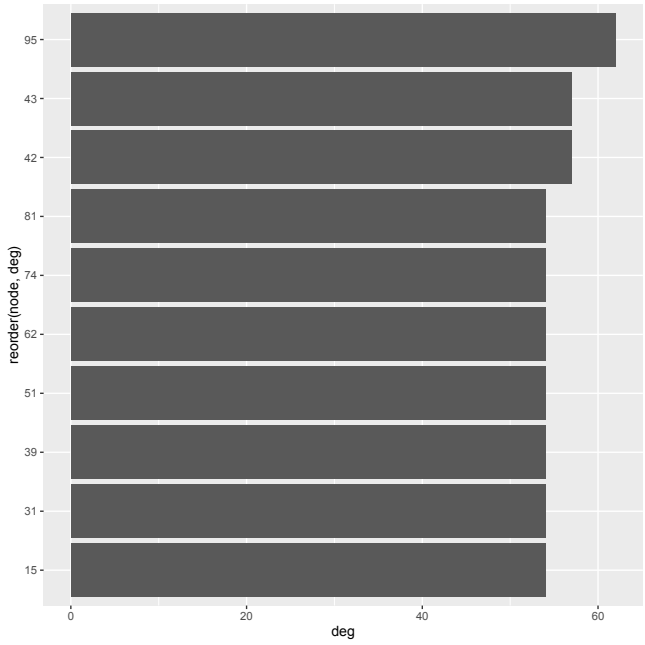


Fig. 6. Top 10 positions in the number of degrees

b) Link Analysis

To analyze cohesion, the density of the network was firstly observed, which is the proportion of all ties that can be theoretically present [7]. Density is defined as the number of connections a participant has divided by the total possible connections a participant could have [15].

Besides, connectedness is also used to analyze the connection of this network. The code for this part is shown in Fig. 7 below.

```
library(igraph)
library(sna)
positions <- read.csv("data/sna_positions.csv", header = T)
relations <- read.csv("data/sna_edges.csv", header = T)

# Construct the network
nodes <- data.frame(positions[, 1])
edges <- data.frame(from = relations[, 1], to = relations[, 2])
net <- graph_from_data_frame(edges, vertices = nodes, directed = FALSE)

# Density, connectedness
# edge_density(net)
gden(as.matrix(get.adjacency(net)))
connectedness(as.matrix(get.adjacency(net)))
```

Fig. 7. Code to calculate density and connectedness

The result of this code shows that the density is 0.2854545 and the connectedness is 0.903232.

c) Community Detection

Social networks contain many communities, which are specifically represented by circles with certain common characteristics [1]. The positions are grouped into different communities according to the functions and duties of the positions. In the same community, the positions are closely connected and the behavior preferences are similar. The individual connections between the communities are relatively loose, and their functional orientations are different. Therefore, identifying different communities from social networks is an important way to accurately select positions.

Cluster_louvain, cluster_edge_betweenness and cluster_label_prop methods are used in this research. The source code is Fig. 8 and the network visualization is shown in Fig. 9., Fig. 10 and Fig. 11.

```
# 1) Louvain method for community detection
clv <- cluster_louvain(net)
crossing(clv, net)
plot(clv, net,vertex.label.cex = 0.5, vertex.size = 10)

# 2) Community structure detection based on edge betweenness
ceb <- cluster_edge_betweenness(net)
plot(ceb, net)

# 3) Community detection based on spread labels
clp <- cluster_label_prop(net)
plot(clp, net)
```

Fig. 8. Network visualization of the positions

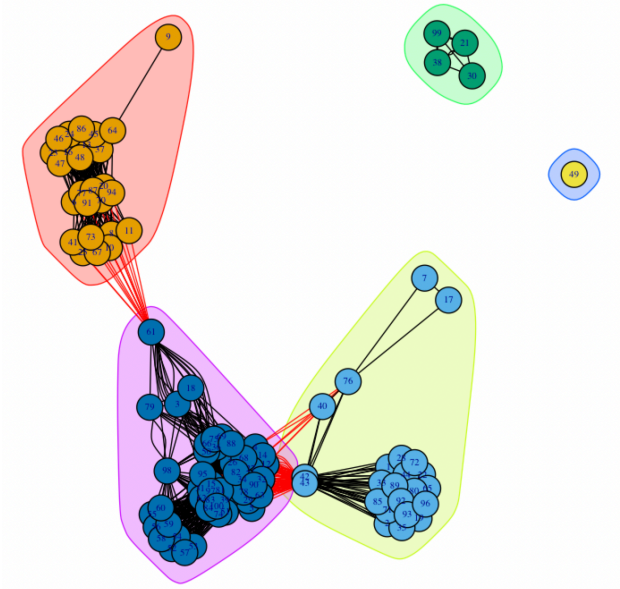


Fig. 9. Community structure by multi-level optimization of modularity

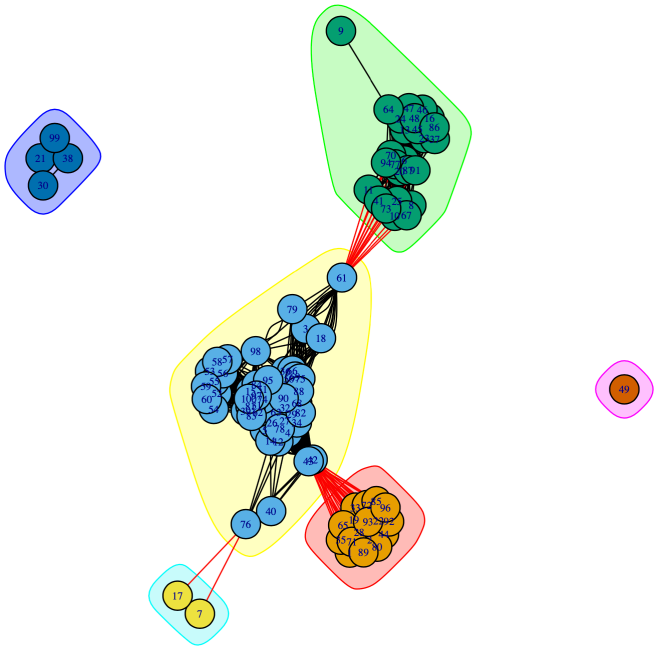


Fig. 10. Community detection based on edge betweenness

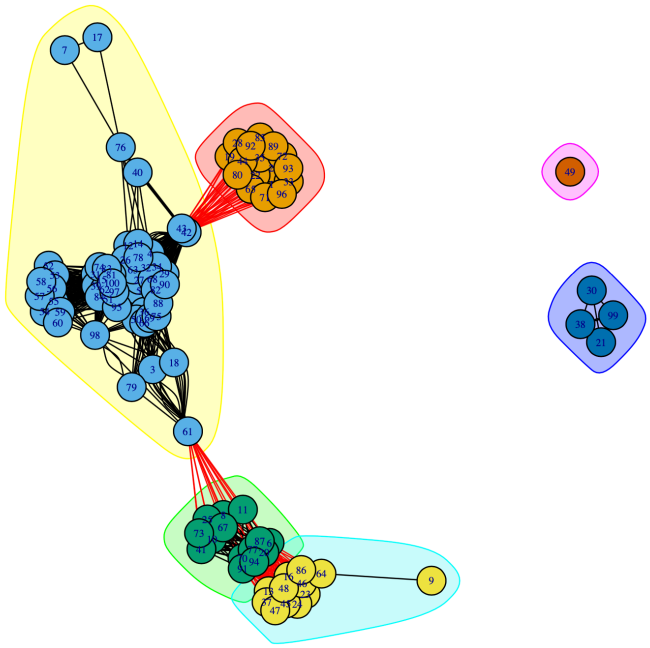


Fig. 11. Community structure detection based on spread labels

d) Graph Cluster Analysis

Graph cluster analysis provides a macro-level view of the dataset. The task of graph cluster is to partition a graph into natural groups so that the nodes in the same cluster are closer to each other than to those in other clusters. There are many algorithms for graph clustering, like k-Spanning Tree, Shared Nearest Neighbor, Betweenness Centrality Based, Kernel k-means, Maximal Clique Enumeration and Highly Connected Components [4].

In this research, the HCS Clustering algorithm is used which is an algorithm based on graph connectivity for cluster analysis. It works by representing the similarity data in a similarity graph and then finding all the highly connected subgraphs. It does not make any prior assumptions on the number of clusters. Kernel k-means is an extension of the standard k-means clustering algorithm that identifies nonlinearly separable clusters. It is used to generate the results for analyzing compared with the HCS Clustering algorithm.

The execution code of these two algorithms is as follows in Fig. 12.

```
library(igraph)
library(RBGL)

positions <- read.csv("data/sna_positions.csv", header = T)
relations <- read.csv("data/sna_edges.csv", header = T)

# Construct the network
nodes <- data.frame(positions[, 1])
edges <- data.frame(from = relations[, 1], to = relations[, 2])
net <- graph_from_data_frame(edges, vertices = nodes, directed = FALSE)

# Graph cluster by HCS
source("algorithm/HCSClustering.R")
HCSClustering(net, kappa = 2)

# Graph cluster by k-means
library("kernlab")
lapKern = laplacedot(sigma = 1)
adj <- as.matrix(get.adjacency(net))
K = kernelMatrix(lapKern, adj)
kmeans(K, 3)
```

Fig. 12. Code to cluster graph

e) Frequent Terms Analysis

Furthermore, the most frequent skills are analyzed in the professional description of the positions. To effectively study the skills required for the position, 4,367 numbers of data crawled so far has been used. The code reads the position description in all job information and then splits the word in the description. However, the frequency of some meaningless stop-words appears relatively high, which will affect the final word frequency [6]. Then I got the stop-words list from the Internet and saved them in the CSV file. When writing the code, I removed stop-words to get the final word frequency data, and sorted by the frequency of occurrence. Finally, the skills in the position are analyzed.

The source code for generating the word cloud is shown in Appendix 6. The following TABLE II is the top 20 skills with the highest frequency. These data are drawn into a word cloud, the larger the word in the word cloud, the higher the frequency of occurrence, so we can clearly notice the required skills.

TABLE II. FREQUENT TERMS

Word	*Counts*	*Word*	*Counts*
experience	15607	knowledge	5355
network	14261	development	5056
data	9760	systems	5019
analysis	9735	skills	5007
work	8261	management	4824
security	6814	working	4193
support	6648	software	4177
intelligence	6170	technical	4146
team	6037	business	3898
social	5823	tools	3767

IV. Analysis of Results and Discussion

A. Structural Analysis

Measuring the influence of individual nodes in the network is one of the focuses of social network analysis. Measuring node centrality helps determine which individuals are more important than others, and identify influential core nodes in the network.

TABLE III. CENTRALITIES

Position ID	*Position Title*	*Degree*	*Closeness*	*Betweenness*
95	Data Scientist	62	0.1504559	293.5095013
43	Senior Digital Network Intelligence Analyst	57	0.1466667	699.5040145
42	Mid-Level Digital Network Intelligence Analyst	57	0.1466667	699.5040145
81	Data Scientist	54	0.1443149	33.8655216
74	AI Data Scientist	54	0.1443149	33.8655216
62	Scala Big Data Developer	54	0.1443149	33.8655216
51	Data Scientist, Trust & Safety	54	0.1443149	33.8655216
39	Data Scientist, Analytics - Responsible AI	54	0.1443149	33.8655216
31	Hadoop Data Scientist	54	0.1443149	33.8655216
15	Data Scientist	54	0.1443149	33.8655216

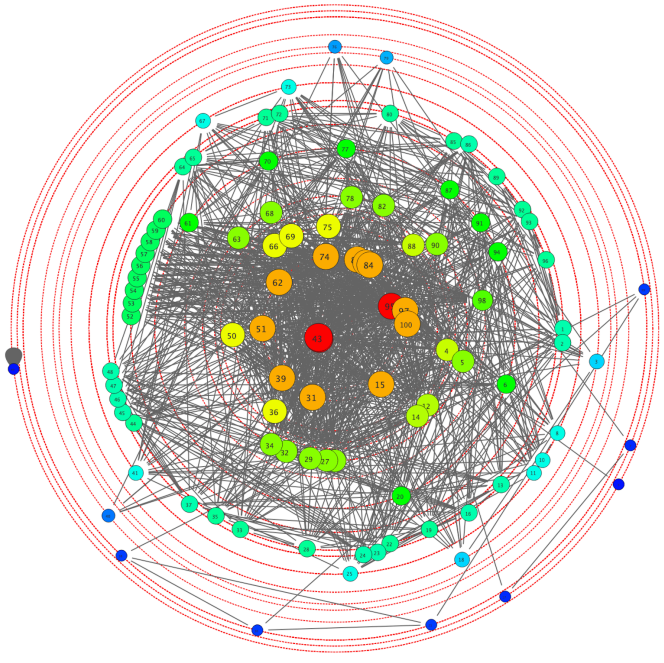


Fig. 13. Network visualization of the positions using SocNetV

Table III shows the specific value of centralities, which shows that the degree of positions with ID 95, 42, 43 is the largest. As shown in Fig. 13 generated by SocNetV, the node closer to the center of the circle indicates that its degree is greater. At the same time, the closeness centrality and betweenness centrality of the three nodes are more than the other nodes which indicate that these three nodes have a very important position and influence in the network. The clearer graph is in Appendix 4 by SocNetV.

Above all, it can be found that different nodes are connected by the function they shared in the position. Different nodes have different degrees and a larger centrality means that it is more important in the network.

B. Link Analysis

Density, connectedness and egocentricity are used to analyze the connection strength of this whole position network.

1) Density: The density of this network is 0.2854545 which shows that the network is relatively tight.
2) Connectedness: The collectedness is 0.903232 which indicates that not all positions relate to each other.
3) Egocentricity: Five positions with ID 95, 43, 42, 81 and 74 are calculated to analyze the connection strength. The code for this process is shown in Fig. 14.

```
ego95 = ego.extract(as.matrix(get.adjacency(net)), 95)
ego43 = ego.extract(as.matrix(get.adjacency(net)), 43)
ego42 = ego.extract(as.matrix(get.adjacency(net)), 42)
ego81 = ego.extract(as.matrix(get.adjacency(net)), 81)
ego74 = ego.extract(as.matrix(get.adjacency(net)), 74)
gden(ego95)
gden(ego43)
gden(ego42)
gden(ego81)
gden(ego74)
connectedness(ego95)
connectedness(ego43)
connectedness(ego42)
connectedness(ego81)
connectedness(ego74)
```

Fig. 14. Code to calculate egocentricity

TABLE IV. ANALYSIS OF SPECIFIED EGO

Position ID	*Position Title*	*Density*	*Connectedness*
95	Data Scientist	0.8415459	1
43	Senior Digital Network Intelligence Analyst	0.6197388	1
42	Mid-Level Digital Network Intelligence Analyst	0.6197388	1
81	Data Scientist	0.8911205	1
74	AI Data Scientist	0.8911205	1

As is shown in TABLE IV, by calculating the density and connectedness, I found that the position of 81 and 74 have the highest density of these five nodes. Higher density means that this node has more links with other nodes.

Combining the calculation results of Table III and Table IV, I found that ID 95, 81, 74 are the most linked and most influential in the whole network. Therefore, RQ I can be answered that positions with ID 95, 81 and 74 are the core and valuable positions.

C. Community Detection

The purpose of community identification is to find the inherent community structure of the network by dividing the nodes into several node groups according to the edge

relationship between the nodes. According to the number of nodes in the network that belong to the community, the community can be roughly divided into overlapping communities and non-overlapping communities [3]. The overlapping community refers to some nodes in the network that not only belong to a community but may also belong to other communities.

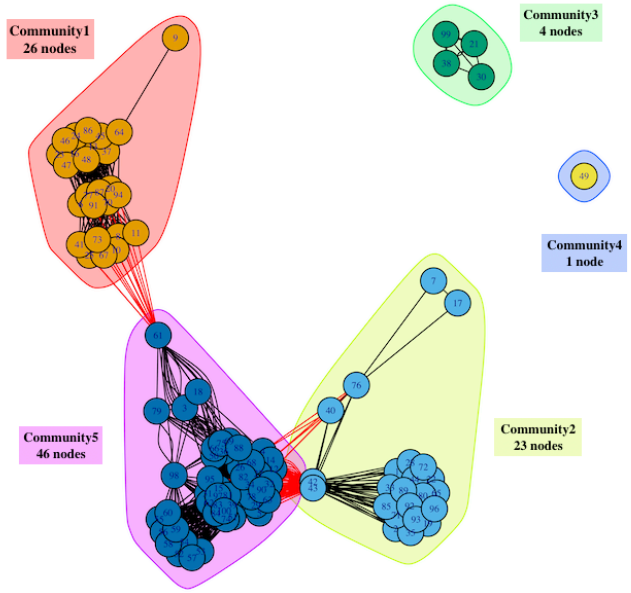


Fig. 15. Visualization of communities after marking community ID and nodes

After marking the communities and counting the nodes, the community distribution map is displayed in Fig. 15 and Appendix 5. The network is divided into 5 communities, and the largest community named Community 5 contains 46 nodes, which means that 46% of the positions belong to this community. Community1 and Community2 contain more than 20% positions, but Community3 has only 4 positions. Community4 is isolated which means that other positions do not have the same function as it.

For RQ II, by combining social network data and comparing Fig. 4 and Fig. 15, we can find:

(1) After the complex network structure in Fig. 4 is divided into communities, the social network structure is clearer. Besides, the internal relations of the communities are relatively close, while the connections between the communities are relatively sparse.

(2) It tends to cluster the nodes with the largest degree in with clear boundaries. In other words, positions with large degrees often form their communities, and they are less in the same community.

D. Graph Cluster Analysis

Graph clustering is an important topic, which involves graph clustering. The data of the clustering problem can be represented as a graph, where each element to be clustered is represented as a node, and the distance between the two elements is modeled by a certain weight on the edge of the link node.

Different methods used in cluster analysis often lead to different conclusions. In my research, the HCS Clustering algorithm and the Kernel k-means algorithm are used to compare the results. The results of the clusters are shown in TABLE V and TABLE VI.

TABLE V. RESULT OF HCS CLUSTERING ALGORITHM

Cluster ID	*Number of positions*
1	83
2	2
3	15

TABLE VI. RESULT OF KERNEL K-MEANS

Cluster ID	*Number of positions*
1	4
2	26
3-19	1
20	49
21-22	1
23	2

It shows that there are 3 clusters after the HCS Clustering algorithm. One of them contains more than 80% of the positions. There is also a cluster that contains only two positions.

The Kernel k-means algorithm gets 23 clusters in this research. There are 49 positions in the largest cluster and 26 positions in another cluster. However, there are 19 clusters that contain only one position.

As we know that different algorithms may lead to different clusters. Although the results are different, some nodes are still in a cluster, which shows that the connection between them is very strong, and there are many direct common features. From this point of view, most of the nodes in this graph are clustered together, only a small part is separated, that is to say, most of the positions are similar. The goal of cluster analysis is to obtain groupings or clusters of similar samples. The output of a cluster analysis method is a collection of subsets of the object set termed clusters characterized in some manner by relative internal coherence and/or external isolation, along with a natural stratification of these identified clusters by levels of cohesive intensity [5].

E. Frequent Terms Analysis

Fig. 16. Word cloud of the frequent Terms

RQ III can be explained from the vivid graph in Fig. 15 and Appendix 7, some words are displayed largely as experience, network, data, security, intelligence, etc. They are high-frequency words that appear in job descriptions, and they all represent the skills required for this position. The characterization of the obtained clusters by interpreting their correspondent word cloud can lead us to draw some conclusions about the different professional groups inside the positions.

Inconsistent with the algorithm to be mastered in this position, the results of the word cloud appear as experience, security, team, management, software, business, tools, etc. These words are more about a wide range of skills and show an overview of the skills in this position. In fact, the company mentioned more in the recruitment needs like skills, rather than mastering an algorithm.

This also reminds our job seekers to pay attention to experience, management and business when developing their skills.

V. Conclusion

A. Summary

Having extracted information from the positions available on LinkedIn, this research analyzed the characteristics of the positions and identified the most valuable positions in methods of centrality, density, connectedness, etc. By applying community detection and graph cluster analysis, it is also found that positions are closely connected and the behavior preferences are similar in the same community. The individual connections between the communities are relatively loose, and their functional orientations are different. The high-frequency skills in the position can also be analyzed, which can help job seekers find the skills they need to better improve themselves in their careers.

Through the whole process of data analysis, this research has reached the following contributions:

1) This research provides the code of the crawler on LinkedIn, which can be helpful to other developers for learning.

2) This research has published a dataset of positions on LinkedIn, which can be used for data analysis by other data analysts.

3) This research uses social network analysis techniques to analyze the connections among positions, helping job seekers to find the most valuable positions.

4) This research also analyzes and mines the key skills in the position, which can help job seekers to improve their abilities.

B. Limitations

The result shows that we can find the relationships among positions and identify the most valuable positions. However, there may be some limitations to this research.

When doing social network analysis, 100 of all positions are randomly selected for analysis. If the experiment uses other data or more data, the results of the experiment may be different. Perhaps, it would be convenient in future research to broaden the range of the sample to do more experiments that would strengthen the results.

Although LinkedIn is a representative place for global job recruitment, it does not exclude that some companies are recruiting on other websites. This will lead to the lack of this part of the data, which will affect the experimental results.

C. Future Work

After this research, I will continue to study the social network analysis of positions. My crawler will collect more position data, which can provide more data for analysis.

Besides, I will also analyze the data of positions under other titles and compare the differences with the social network analysis positions in this research.

Future more, I will continue to read a lot of literature, study the most advanced social network analysis techniques and algorithms, and apply them to my research.

Acknowledgment

I am very lucky to have this opportunity to do research on social network analysis. I would like to express my thanks to Prof. Wong for his careful teaching in the online class and patient answering in the Q & A class, as well as providing some learning materials for us. In the future, I will also use the knowledge I have learned to serve the enterprise in the software development and continue studying to improve myself.

References

[1] Fortunato S.Community detection in graphs[J].Physics Reports, 2010, 486(3):75-174.

[2] Lancichinetti, Andrea; Fortunato, Santo (2009-11-30). "Community detection algorithms: A comparative analysis". Physical Review E. American Physical Society (APS). 80 (5).

[3] Sun P G, Gao L, Han S S. Identification of overlapping and non-overlapping community structure by fuzzy clustering in complex networks[J]. Information Sciences, 2011, 181(6): 1060-1071.

[4] Wikipedia contributors. (2020, April 24). Cluster analysis. In Wikipedia, The Free Encyclopedia. Retrieved 06:28, May 25, 2020, from https://en.wikipedia.org/w/index.php?title=Cluster_analysis&oldid=952923740

[5] Matula D W. Graph theoretic techniques for cluster analysis algorithms[M]//Classification and clustering. Academic Press, 1977: 95-129.

[6] A. N. K. Zaman & P. Matsakis , C. Brown (2011) “Evaluation of Stop Word Lists in Text Retrieval Using Latent Semantic Indexing”, In: Sixth International Conference on Digital Information Management (ICDIM), pp. 133-136, Melbourne.

[7] Reamer, F. G. (2013). Social work in a digital age: Ethical and risk management challenges. Social Work, 58(2), 163-172. https://doi.org/10.1093/sw/swt003

[8] Wikipedia contributors. "Louvain modularity." Wikipedia, The Free Encyclopedia. Wikipedia, The Free Encyclopedia, 18 Apr. 2020. Web. 26 May. 2020.

[9] Clauset, Aaron; Newman, M. E. J.; Moore, Cristopher (2004-12-06). "Finding community structure in very large networks". Physical Review E. 70 (6): 066111

[10] Wikipedia contributors. "HCS clustering algorithm." Wikipedia, The Free Encyclopedia. Wikipedia, The Free Encyclopedia, 2 May. 2020. Web. 26 May. 2020.

[11] Social Network Visualizer https://sourceforge.net/projects/socnetv/

[12] Zhang C, Song S M. Forward position analysis of nearly general Stewart platforms[J]. 1994.

[13] Skeels M M, Grudin J. When social networks cross boundaries: a case study of workplace use of facebook and linkedin[C]//Proceedings of the ACM 2009 international conference on Supporting group work. 2009: 95-104.

[14] Zide J, Elman B, Shahani-Denning C. LinkedIn and recruitment: How profiles differ across occupations[J]. Employee Relations, 2014, 36(5): 583-604.

[15] Wikipedia contributors. "Social network analysis." *Wikipedia, The Free Encyclopedia*. Wikipedia, The Free Encyclopedia, 4 May. 2020. Web. 26 May. 2020.

[16] Igraph social network analysis http://rstudio-pubs-static.s3.amazonaws.com/484205_48e8ef6ea31b4bb7a11746c988da39de.html

[17] Zhang X K, Ren J, Song C, et al. Label propagation algorithm for community detection based on node importance and label influence[J]. Physics Letters A, 2017, 381(33): 2691-2698.

[18] Bastian M, Heymann S, Jacomy M. Gephi: an open source software for exploring and manipulating networks[C]//Third international AAAI conference on weblogs and social media. 2009.

[19] Dhillon I S, Guan Y, Kulis B. Kernel k-means: spectral clustering and normalized cuts[C]//Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining. 2004: 551-556.

APPENDICES

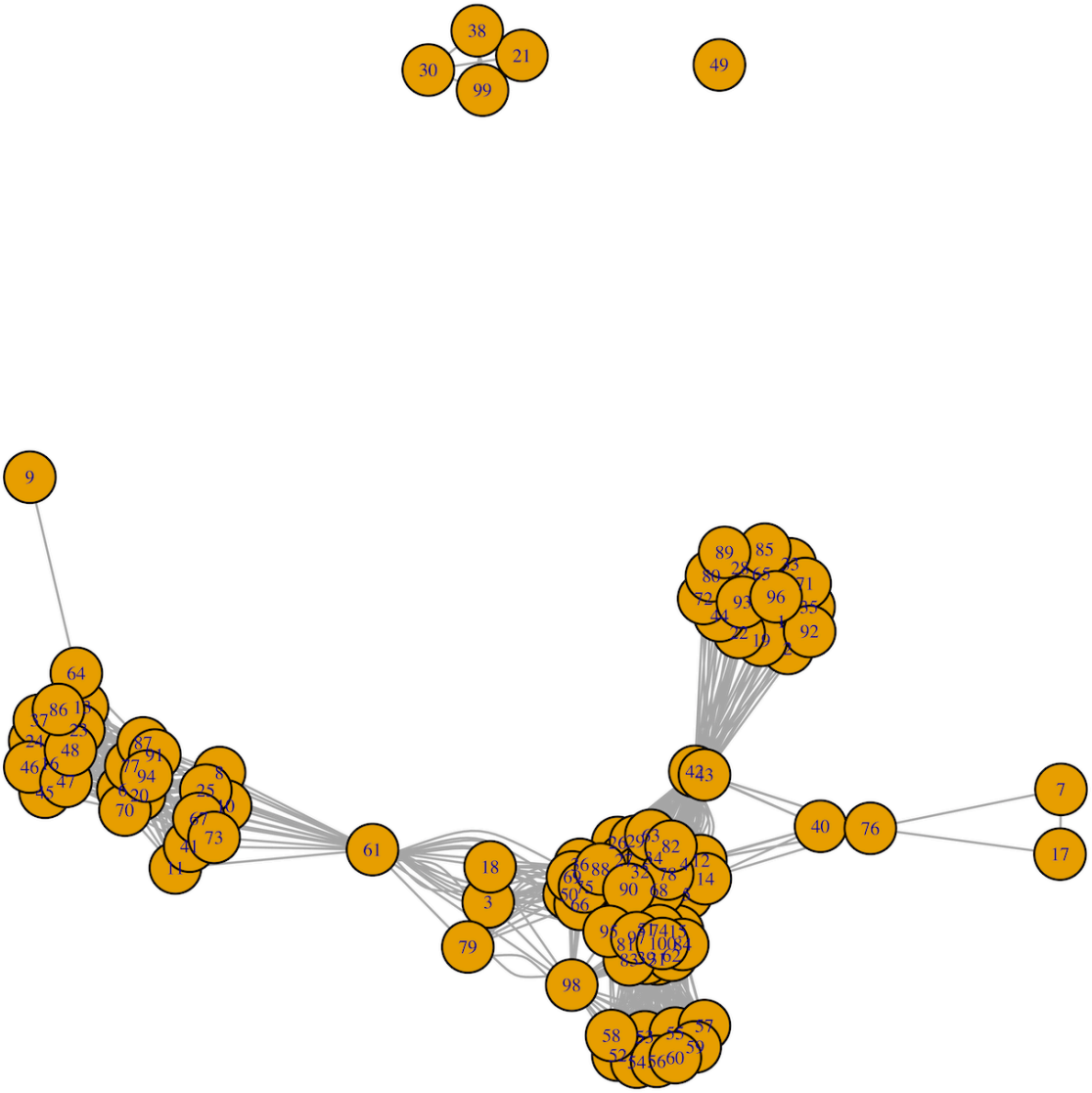
Appendix 1. Example data of nodes

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
id	keywords	jobId	jobTitle	description	industry	jobFunction	companyNa	employmen	comp	companyAd	seniorityLev	pubTime	createdTime	updatedTime	
1	social network analysis	1870208712	DevOps Engineer Cloud P	DevOps Eng	Information Technol	Other	Client Serve	Full-time		London, En	Entry level	2020/5/19	19/5/2020 2	19/5/2020 20:39:43	
2	social network analysis	1793085047	Senior Data Scientist	DescriptionJ	Defense & Space	Other	Reston, VA	Full-time	https:	Reston, VA	Mid-Senior	2020/5/19	19/5/2020 2	19/5/2020 20:39:43	
3	social network analysis	1865976463	Data Analyst	Findasense	Marketing and Adve	Analyst,Research	Covibar Mac	Full-time	https:	Covibar Mac	Associate	2020/5/19	19/5/2020 2	19/5/2020 20:39:44	
4	social network analysis	1789061042	Remote Principal Enginee	JOB DESCRI	Research	Information Technology,F	Kimetrica, LI	Full-time		Washingtor	Mid-Senior	2020/5/19	19/5/2020 2	19/5/2020 20:39:46	
5	social network analysis	1869793420	Contract - Tech Professio	Contract - T	Information Technol	Information Technology	Malibu, CA	Contract	https:	Malibu, CA	Associate	2020/5/19	19/5/2020 2	19/5/2020 20:39:46	
6	social network analysis	1869904971	Crebs - Creative and Tech	In previsione di un ampliamento		Marketing,Sales	Reallife Tele	Full-time		Rome, Latiu	Entry level	2020/5/19	19/5/2020 2	19/5/2020 20:39:47	
7	social network analysis	1869967101	Learning and developmer	REPORT TO	Staffing and Recruiti	Human Resources	Montreal, Q	Full-time	https:	Montreal, Q	Entry level	2020/5/19	19/5/2020 2	19/5/2020 20:39:49	
8	social network analysis	1868847104	Internship for Social Medi	Job Respons	Marketing and Adve	Marketing,Public Relation	SOCIO Intell	Internship		Federal Terr	Internship	2020/5/18	19/5/2020 2	19/5/2020 20:39:50	
9	social network analysis	1844635875	Fiduciary Advisor II	Position Ov	Financial Services	Finance	Blue Bell, PA	Full-time	https:	Blue Bell, PA	Not Applica	2020/5/18	19/5/2020 2	19/5/2020 20:39:51	
10	social network analysis	1869580663	Social Media Manager - F	In previsione di un ampliamento		Marketing,Public Relation	Reallife Tele	Full-time		Rome, Latiu	Associate	2020/5/18	19/5/2020 2	19/5/2020 20:39:52	
11	social network analysis	1822841450	Stage Social Media Marke	Feat Food a	Food & Beverages	Marketing	Segrate, Lor	Internship	https:	Segrate, Lor	Internship	2020/5/19	19/5/2020 2	19/5/2020 20:39:53	
12	social network analysis	1870212407	Project Manager	An exciting	Higher Education	Project Management,Info	Peasedown	Full-time	https:	Peasedown	Entry level	2020/5/18	19/5/2020 2	19/5/2020 20:39:53	
13	social network analysis	1870357408	Social Media Account Mar	Social Media	Information Technol	Sales,Business Developme	Ggruk	Full-time		London, En	Entry level	2020/5/18	19/5/2020 2	19/5/2020 20:39:55	
14	social network analysis	1826060525	Manager, BIE Global Fulfil	Description/	Computer Software	Project Management,Info	Bellevue, W	Full-time	https:	Bellevue, W	Mid-Senior	2020/5/18	19/5/2020 2	19/5/2020 20:39:56	
15	social network analysis	1790269358	Data Scientist	DescriptionJ	Defense & Space	Engineering,Information `	McLean, VA	Full-time	https:	McLean, VA	Mid-Senior	2020/5/18	19/5/2020 2	19/5/2020 20:39:58	
16	social network analysis	1868346491	Irregular Warfare Analysis	Responsibili	Information Technol	Business Development,Sa	Reston, VA	Contract	https:	Reston, VA	Associate	2020/5/18	19/5/2020 2	19/5/2020 20:39:59	
17	social network analysis	1836554253	Director, Regional Recruit	Job Summa	Hospital & Health Ca	Human Resources	Accord Care	Full-time		Marietta, GA	Director	2020/5/18	19/5/2020 2	19/5/2020 20:40:01	
18	social network analysis	1868279879	Dir of Analytics	Our Client is	Hospital & Health Ca	Research,Strategy/Plannir	Healthcare A	Full-time		Washingtor	Director	2020/5/18	19/5/2020 2	19/5/2020 20:40:02	
19	social network analysis	1868334086	Senior Artificial Intelligen	Artificial Int	Information Technol	Other	Arlington, V	Full-time	https:	Arlington, V	Associate	2020/5/18	19/5/2020 2	19/5/2020 20:40:04	
20	social network analysis	1864425891	Digital Marketing Speciali	At Maven D	Hospital & Health Ca	Marketing,Sales	Gold Coast,	Contract	https:	Gold Coast,	Associate	2020/5/18	19/5/2020 2	19/5/2020 20:40:04	

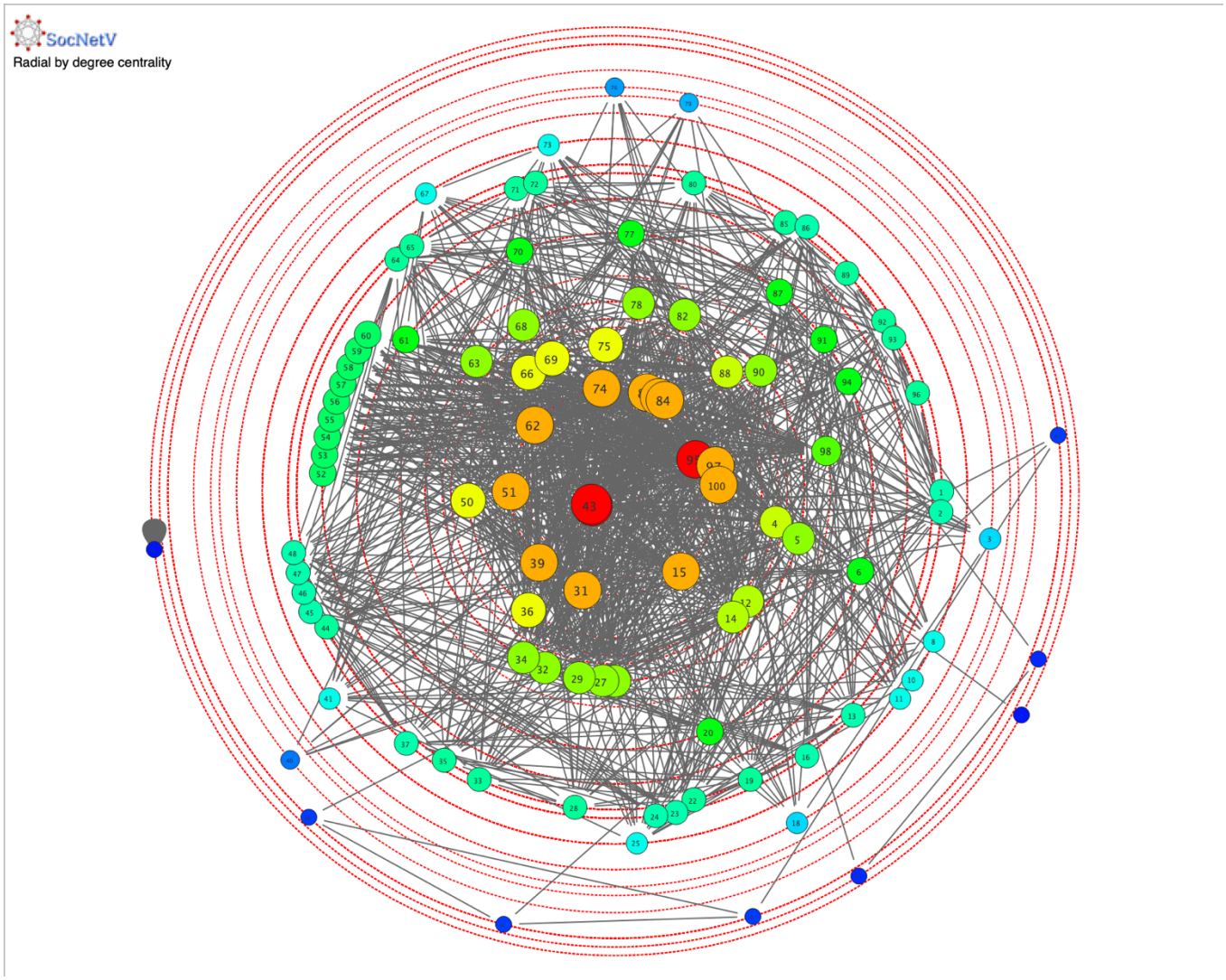
Appendix 2. Example data of edges

A	B	C
fromId	toId	jobFunction
1	2	Other
1	19	Other
1	22	Other
1	28	Other
1	33	Other
1	35	Other
1	42	Other
1	43	Other
1	44	Other
1	65	Other
1	71	Other
1	72	Other
1	80	Other
1	85	Other
1	89	Other
1	92	Other
1	93	Other
1	96	Other
2	19	Other

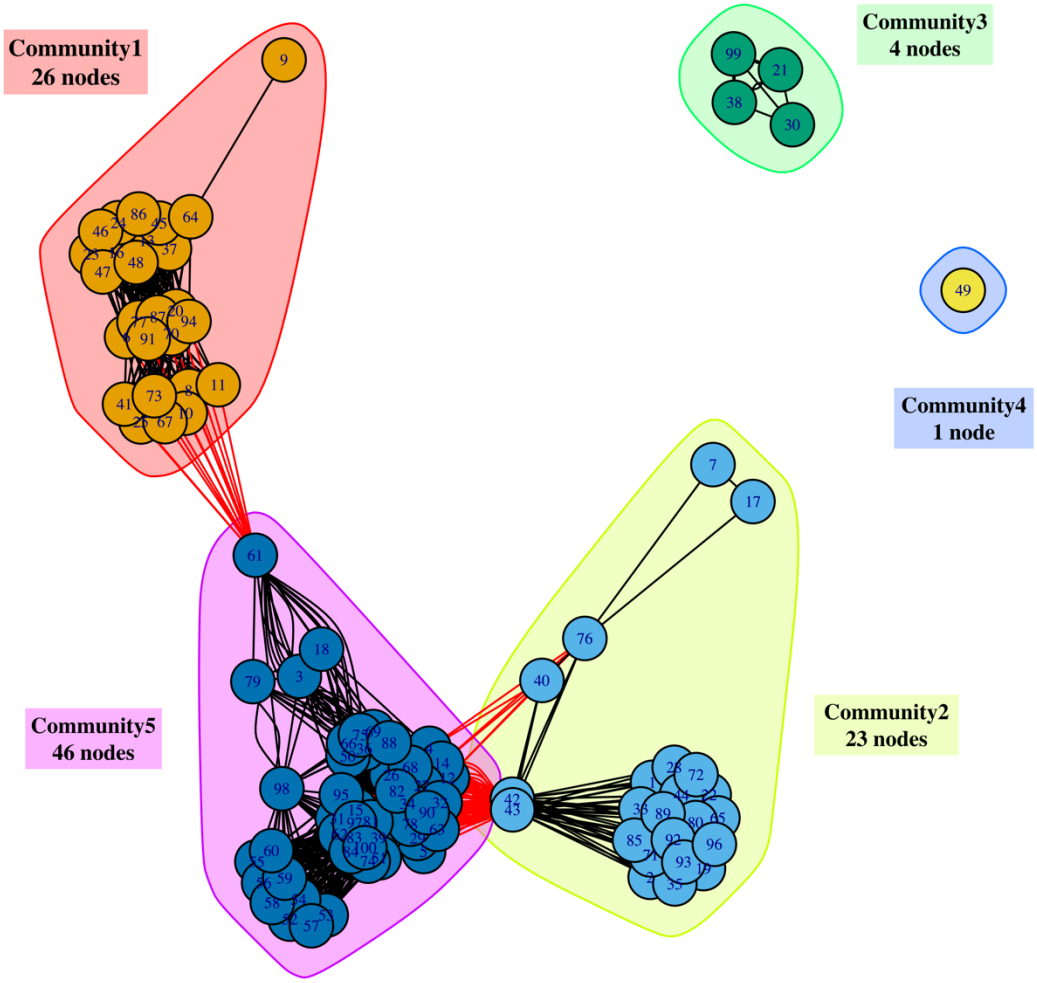
Appendix 3. Network visualization of the positions



Appendix 4. Network visualization of the positions using SocNetV



Appendix 5. Visualization of communities after marking community ID and nodes



Appendix 6. Code to generate the word cloud

```
library(wordcloud2)
library(dplyr)
library(textfeatures)

# Read the position description
description = readLines('data/position_description.txt')
head(description)

txt = description[description != ""]
txt = tolower(txt)
txtList = lapply(txt, strsplit, " ")
txtChar = unlist(txtList)
# clean symbol(.,!:;?)
txtChar = gsub("\\.|,|\\!|:|;|\\?", "", txtChar)
txtChar = txtChar[txtChar != ""]
data = as.data.frame(table(txtChar))
colnames(data) = c("Word", "freq")
ordFreq = data[order(data$freq,decreasing=T),]

# Filter the stopwords
df = read.csv('data/stopwords.csv', header = T)
Word = select(df,Word)
antiWord = data.frame(Word, stringsAsFactors = F)
# ordFreq - antiWord
result = anti_join(ordFreq, antiWord, by = "Word") %>% arrange(desc(freq))

result = result[1:50,]
head(result, 20)

# Draw graph
wordcloud2(data=result, size=1)
```

Appendix 7. Word cloud of the frequent Terms