Analysis of Researchers’ Internal Working Network

Ying Qin1, Qing Zhang1, CuiFang Zheng1, ZhiYu Tian1, Yong Wen1, XinKun Chu1, Hao Zhang1, YingPing Zhou2 and Hao Zhang1*

1Institute of Computer Application, China Academy of Engineering Physics. Mianyang, Sichuan, 621900, China
2Mianyang first Military Representative Office, Mianyang, Sichuan, 621900, China
*Corresponding author’s e-mail: linusec@163.com

Abstract. In this paper, we adopt social network analysis technology to construct the working network of the scientific research personnel, working in the isolated network environment of a military research institute, based on the integrated edge data obtained by processing three groups of work records including emails, meetings and travels, and then we complete the calculations of indicators including Degree centrality, Closeness centrality, Betweenness centrality and Eigenvector centrality. Finally, we evaluate the performance of scientific research personnel, explore the working relationship between them, and find out who is working abnormally, which provides a new method for staff appraisal.

1. Introduction

Researchers in military research institutes need to perform daily tasks in an isolated network environment due to special work requirements, and a large number of work records are generated during the working, such as email records, meeting records, travel records and so on. These records contain a large number of information about employee work dynamics, as well as working connections between them, which forms a huge internal working network. In this network, it is worth exploring to find out where each employee is and what role each of employee plays. This provides a new way of investigating the employee's working status, observing the working relationship between the employees in the institutes, and even finding whose work is abnormal, which has great practical value.

Social network analysis (SNA) technology is a fashionable network analysis technology, and has produced a large number of practical applications in various network analysis cases, such as human resource networks. Literature [1] uses social network analysis technology as a unique method to study the importance of social relationships in human resource development (HRD), which is used to measure the impact of interpersonal relationships on human capital, and it is proposed that SNA can improve the empirical rigor of HRD theory building in such areas as organizational development, organizational learning, leadership development, organizational change, and training and development; the literature [2] describes how SNA technology help HRD departments making analysis and intervene in knowledge management through several cases, such as case 1 by constructing a social network of workers, the analysis found that there was a gap in knowledge sharing between the company groups, which hindered information flow and knowledge innovation, and took measures such as arrange liaisons, arrange all teams to work in the same room and redistribute decision-making power; in case 2, also through SNA, it is found that a large company's cross-regional connection relies on only a few high-level brokers, which brings risks to the company, so they make a change to
improve this problem through incentive mechanism, regular cross-regional meetings, establishment of contact points and publication of succession plans; the literature [3] use SNA technology to build company's social network, calculate various indicators of social networks, and then use data mining techniques to construct a laid off predicting model based on decision tree, which accuracy is above 86%.

Based on the analysis of the 2018 email, meeting and travel records (see Table 1) from some research institute, this paper extracts three groups of edge data, and combines these three groups of edge data into one comprehensive edge data, which is then used to construct the working network of the institute using SNA technology, and the network indicators are calculated and analyzed. It is verified that the working network constructed by this method is closer to the actuality, which can reflect the working relationship between employees of the institute, and provides data on how to evaluate the performance of employees in the working network and find whose working status is abnormal.

| Table 1. Records Summary |
|--------------------------|
| **Email records** | **Meeting records** | **Travel records** |
| Count | 142902 | 130320 | 3449 |
| Time range | 2018 | 2018 | 2018 |
| Features | Sender, Recipient, Time, Mail label, Mail location | Participants, Departments, Time, Location | Traveler, Time, Location |

2. Methodology
The flow of algorithm includes preprocessing the data of emails, meetings and travels respectively, counting the three groups of edge data, and then obtaining the integrated edge data, using the ForceAtlas2 network topology algorithm and the Louvain community discovery algorithm to build working network and calculate the network indicators (see Figure 1).
2.1. Data preprocessing
Data preprocessing refers to removing the redundant information of the emails, meetings and travels respectively, aiming for making the data available for obtaining edge data. The specific pre-processing steps are shown in the figures below.

![Figure 2. Email Records Preprocessing](image)

![Figure 3. Travel Records Preprocessing](image)

![Figure 4. Meeting Records Preprocessing](image)

2.2. Edge Data Counting
Edge data is the basis for building a working network [4][5], it can be simply expressed as the number of work activities between employees, such as the number of emails, the number of times they attend same meeting, or the number of being on same travel. Generally, the more edge data of the employee or the larger the value of the edge data, the better the evaluation of the employee in the network. However, not all records directly reflect the working relationship, such as attending the same meeting or same travel do not always mean attendees have work relationship between them. If you do not distinguish these situations and counting them in edge data, it will lead into unnecessary noise information and directly affect the subsequent network topology and indicator calculation. In order to mine the exact working relationship, these records need to be weighted at first: records with a high degree of relevance to the work business are assigned a high weight, and vice versa.

Due to the huge amount of work records, if you assign weight to them one by one, the labor cost will be unbearable. This paper adopts another way to assign weights based on the features of the three groups of data.

a) Email Edge Data
The email weight is assigned based on the email type and the number of email recipients, as \( W_s \) and \( W_n \), and then multiply the two number to get the final weight \( W_{email} \).

\[
W_{email} = W_s \times W_n \quad (1)
\]

The email is first weighted according to the type of mail (Table 2). There are two groups of email types. One group is to divide the email according to the email relationship into the original email, replied email, forwarded email and CC email. The other group is to divide the email into the inbox
email, personal folder email and trash box email. In this paper, the weights \( W_{ti} \) (i=0,1,2,3) and \( W_{pj} \) (j=0,1,2) are respectively assigned according to the two classification groups, and then the weights \( W_s \) are obtained by multiplying these two weights (see Equation 2). We believe that the correlation between business and various types of email has a relationship that replied email > original email > forwarded email > CC email and personal folder email > inbox email > trash box email, so the weight assignment should also conform to this relationship. The weight values used in this calculation are shown in Table 2. The weights can be appropriately modified according to the specific situation.

Table 2. Assignment of Email Weight \( W_s \)

| Type          | Weight | Value |
|---------------|--------|-------|
| Original      | \( W_{t0} \) | 1     |
| Replied       | \( W_{t1} \) | 3     |
| Forwarded     | \( W_{t2} \) | 0.5   |
| CC            | \( W_{t3} \) | 0.1   |
| Personal Box  | \( W_{p0} \) | 7     |
| Inbox         | \( W_{p1} \) | 1     |
| Trash Box     | \( W_{p2} \) | 0.02  |

\[ W_s = W_{ti} \ast W_{pj} \quad (i = 0,1 \ldots 3; j = 0,1,2) \]  

Secondly, according to the number \( n \) of recipients of the email, the weight distribution \( W_n \) is \( 1/n \). and the final email weight is:

\[ W_{email} = W_{ti} \ast W_{pj} \ast \frac{1}{n} \]  

After the weight assignment is completed, the statistics of the edge data are performed. The statistical methods are divided into directed and undirected: the directed statistical method needs to separately calculate the edge data of employee A to employee B and employee B to employee A; while the undirected statistical method does not distinguish the direction, the statistics of employee A to employee B and employee B to employee A are counted together.

The email edge data uses the directed statistical method. As in the Equation 4, the email edge data of the employee A to the employee B (\( Edge_{email,AB} \)), the count of the emails sent by the employee A to the employee B, is the sum of the weights of all the emails sent by the employee A to the employee B.

\[ Edge_{email,AB} = \sum W_{mail} = \sum W_{ti} \ast W_{pj} \ast \frac{1}{n} \]  

b) Meeting Edge Data
The weight of the meeting is assigned according to the size \( n \) of the meeting, and the value is \( 1/n \). Edge data is counted using undirected statistics.

c) Travel Edge Data
Travel records has no weight assignment, and the edge data of travels between employees is directly counted using undirected statistics.

d) Integrated Edge Data
Through the comparison of the three groups of edge data, we found that for the email data, the contact direction between employees is obvious and can indicate the working connection between the them explicitly, but some employees, due to job requirements (such as the office secretary) or
individual habit, will generate a large number of "noise" emails (such as mass notification emails, life information emails, etc.), which will result in network indicators are higher than the actual situation. On the other hand, meeting and travel records can accurately reflect the work relationship with business, but there is relatively weaker contact and direction relationship between employees. For example, there is not necessarily a working relationship between employees who attend a same meeting or on a same travel (see Table 3).

| Type   | Feature                          | Direction |
|--------|----------------------------------|-----------|
| Email  | Obvious direction                | directed  |
| Meeting| No direction, Strong relation of business | undirected |
| Travel | Smaller amount, No direction, Strong relation of business | undirected |

Table 3. Three Groups of Edge Data Summary

The three groups of edge data complement each other and reflect more completed work characteristics of the employees. For example, there are relatively more meetings and travels for leadership staff; service staff have more emails, but fewer meetings and travels; general staff have fewer emails, meetings, and travels. Therefore, these three groups of edge data should be combined in a way to obtain integrated one that more fully reflects the employee's complete working information. The method adopted in this paper is to assign weights to these three groups of edge data again and sum them to obtain integrated edge data. As in Equation 5, $Edge_{total, AB}$ is the integrated edge data of employee A to employee B. $Edge_{email, AB}$, $Edge_{meeting, AB}$, and $Edge_{errand, AB}$ represent three sets of edge data, $W_{EMAIL}$, $W_{MEETING}$, and $W_{ERRAND}$ represent weights, and the weights' relationship is $W_{ERRAND} > W_{MEETING} > W_{EMAIL}$. The weight values used in this calculation are 7, 5, and 1 respectively.

$$
Edge_{total, AB} = Edge_{email, AB} * W_{EMAIL} + Edge_{errand, AB} * W_{ERRAND} + Edge_{meeting, AB} * W_{MEETING}
$$

(5)

It should be noted that, due to the different statistics way of email edge data, meeting edge data and travel edge data (the email edge data is directed statistics, meeting edge data and travel edge data are undirected statistics), it is necessary to change meeting edge data and travel edge data from the undirected statistics to the directed statistics before merging the three groups of edge data, in order to retain the directed information as much as possible. The method adopted in this paper is to divide the value of the undirected edge data into two, which are used as edge data of the two directions of the directed statistics. For example, in the undirected edge data, there is only the value $m$ of the edge data of employee A and employee B, which is converted into directed edge data, and the data of employee A to employee B is $m/2$, and the data of employee B to employee A is also $m/2$.

2.3. Network Indicators

According to the definition of each vertex indicator in the social network analysis algorithm [4][5][6], the degree centrality, closeness centrality, betweenness centrality and eigenvector centrality are used to evaluate the performance of employees. Their definitions are as follows:

1. Degree centrality, which counts how many vertices have direct working contact with the current vertex, can directly reflect how many people have worked with the employee.

2. Closeness centrality, which calculates how close the vertices are to the center of the entire working network, can reflect whether the employee is marginalized.
3. Betweenness centrality, which calculates whether the vertex is a key connection point of two or more groups of vertexes, usually used to discover key contacts between different groups.

4. Eigenvector centrality, which calculates the influence of vertices throughout the network, often used to discover the true leadership of the entire organization.

3. Results and Discussion

3.1. Working Network Analysis

a) Research Institute's Working Network

Figure 5 is the final working network figure of a research institute's 2018 integrated edge data. The vertices in the figure represent employees. The line between the vertices indicates the work contact and the vertices’ size indicate the value of their own indicator. Different colors are used to distinguish different groups. After observation, you can find the following points from the diagram.

1. The employees of office department of the institute are almost located in the center of the network, while the employees of the other departments are located at the periphery, and the various color groups divided by the algorithm correspond to the actual departments.

2. Bigger vertices are almost the leaders of the institute and department, indicating that they are the main peoples in the working contact activities.

3. The vertices inside the groups of the working network are closer than those between groups, indicating that the working activities in the group is more frequent than between groups.

b) Departments' Working Network

After extracting the edge data of the different departments, and constructing the department's working network, horizontal comparisons of these networks can also find some useful information for departments. In Figure 6 (a), employees from high-level to low-level are hierarchically arranged from right to left, indicating that the department's management is a layer-by-layer mode, which is a typical network structure of large departments; In Figure 6 (b), the distinction between employees is relatively small, mostly belongs to smaller departments' structure, the employees are closely connected and have full work arrangement. Some departments are shown like Figure 6(c), whose network is divergent. High-level employees in the center are too big, but the peripheral vertices are too small, indicating that the low-level employees have weak working relationships. Most departments like the structure of
Figures 6(a) and 6(b). The department of Figure 6(c) has been verified that there is indeed a situation that the staff is loose and can do more work.

3.2. Network Indicators Analysis

a) Employee Indicators Analysis

Using SNA technology, each employee's indicators (degree centrality, closeness centrality, betweenness centrality, and eigenvector centrality) can be calculated and ranked. Through these rankings, you will have a clear understanding of each employee's performance in the institute. For example, through the degree centrality ranking, you can observe who usually contact more people; through the closeness centrality ranking, you can find out who is ranked at the bottom and is marginalized; through the betweenness centrality ranking, you will find out who is the key connection; through the eigenvector centrality ranking, you can observe who has played a leading role.

| Ranking | Degree Centrality | Closeness Centrality | Betweenness Centrality | Eigenvector Centrality |
|---------|-------------------|-----------------------|------------------------|------------------------|
| 1       | Employee A        | Employee C            | Employee E             | Employee G             |
| 2       | Employee B        | Employee D            | Employee F             | Employee F             |
| ...     | ...               | ...                   | ...                    | ...                    |

In addition, after grouping employees according to different levels (leader/middle leader, backbone, general employee), use Kmeans algorithm for cluster analysis and make comparison, the following results are obtained:

| Team | General employee group | Backbone group | Leader/middle level leader group |
|------|------------------------|----------------|---------------------------------|
| 1    | 31                     | 8              | 0                               |
| 2    | 28                     | 10             | 0                               |
| 3    | 26                     | 0              | 0                               |
| 4    | 18                     | 16             | 0                               |
| 5    | 6                      | 17             | 11                              |
| 6    | 0                      | 11             | 18                              |
As can be seen from the Table 5, most of the general employees group are assigned to Team 1, 2 and 3, and the leadership and backbone groups are mostly in Team 5 and 6. In the general employees group, 6 people in Team 5 are grouped with the other two groups (backbone group and leader group) with outstanding performance and they can be considered as next leaders; in the backbone group, the 8 people in the first team was grouped with the majority of the general employees group, and their performance was slightly worse.

**b) Departments Indicators Analysis**

Figure 7 shows the departments’ rankings of the four indicators (ranked by the average indicators value of employees for each of departments). In these rankings, the organs are in the first echelon in all performances, and the other departments are in the second echelon. In the first echelon, the institute leaders are in an absolute position, which indicating that they are the focus of all work activities; in the second echelon, the department B, E and F have outstanding performance, almost in the top three of the second echelon, while the department C and D stand in the last two positions. As a result, adjustments may be required in organizational management and work assignments for these two departments.

It is worth noting that in the degree centrality ranking, the institute leaders ranks first, which indicates that the number of persons that leaders directly contacted is the most, which is necessary for further discussion of whether this is good for the management of the institute; In the betweenness centrality ranking, the Finance Department and the Human Resources Division are in the top two, which is in sharp contrast with the performance in the other rankings, that indicates these two departments usually have more contact with other departments, but their work are not main business. This is in line with the actual situation; in the eigenvector centrality ranking, the institute leaders are leading ahead. This aspect proves the effectiveness of the algorithm in discovering network leaders. On the other hand, this method can be used as an evaluate indicator of mining the leader person.

![Degree Centrality Ranking](image-url)
(b) Closeness Centrality Ranking

c) Betweenness Centrality Ranking
During the analysis, we also found some suspected anomalies. Some employees are in a very remote zone in the work network figure, indicating these employees usually have few working contacts; some employees are far away from their own teams or departments, while have more contacts with other groups (see Figure 8). This should arouse the attention of the relevant departments and analyze whether these situations are normal and whether it is necessary to adjust the work content or jobs of them.
3.3. Future Work

First, the data source should be more comprehensive. In the actual work process, the contact data of the researchers also includes other data such as telephone records, social software records, etc., which are also part of the employee's working activities, and due to each employee's work habits are not completely consistent (some often send Mail, some use the phone as the main contact method), we should count these data in the integrated edge data as much as possible, so that we can get a more accurate work network map and indicator evaluation.

Second, there is still more work to do for improvement in the statistical methods of edge data, such as the "synthesis" algorithm of integrated edge data and the effect of the difference in the total number of data between the three sets of edge data on the results.

Third, when building a working network diagram, the stability of the network diagram is not good enough. In the results of multiple calculations, the position of some employees is greatly deviated or even can’t stabilized. This aspect may be due to the inherent relationship of the data itself, otherwise may also be caused by the visualization algorithm. For this problem, consider further optimizing the ForceAtlas2 algorithm or using other visualization algorithms.

4. Conclusion

After pre-processing the data of three kinds of work activities of mail, conference and business trips of our researchers, this paper obtains comprehensive edge data, uses social network analysis technology, constructs the working network map of scientific research personnel, and quantifies the indicators and mines the work network. The working relationship between the scientific research personnel in the figure, the method can reflect the working relationship between the internal employees of the unit more realistically, and provides a new visual way and basis for the relevant departments to understand the working status between the personnel of the unit, and optimizes the position. Personnel management problems such as configuration and discovery of work abnormalities provide a new way of investigation.

References

[1] John-Paul Hatala. (2006) Social Network Analysis in Human Resource Development: A New Methodology. Human Resource Development Review. J. 5(1): 45-71.

[2] Salvatore Parise. (2007) Knowledge Management and Human Resource Development: An Application in Social Network Analysis Methods. Advances in Developing Human Resources. J. 9: 359.

[3] Hui-Ju Wu, I-Hsien Ting, and Huo-Tsan Chang. (2010) Integrating SNA and DM technology into HR Practice and Research: Layoff Prediction Model. Mining and Analyzing Social Networks. J. Sci 288. 53-66.

[4] Maksim Tsvetovat, Alexander Kouznetsov. (2011) Social Network Analysis for Startups. O'Reilly Media, Boston.

[5] John Scott. (2000) Social Network Analysis: A Handbook. SAGE Publications, London.

[6] Stanley Wasserman, Katherine Faust. (1995) Social Network Analysis: Methods and Applications. Cambridge University Press, Cambridge.