Social Network Analysis - Report

Noemi Benci, Federico Pirona

University of Florence

April 5, 2020

Introduction

The network we are focusing is built on a dataset, used also by Cross and Parker[1], which contains four intra-organizational networks. In particular such network, which is from a research team in a manufacturing company, is based on the employees *"awareness of each others" knowledge and skills* ("I understand this person has knowledge and skills. This does not necessarily mean that I have these skills or am knowledgeable in these domains but that I understand what skills this person has and domains they are knowledgeable in"). The weight scale in this network is: 0: I Do Not Know This Person/I Have Never Met this Person; 1: Strongly Disagree; 2: Disagree; 3: Somewhat Disagree; 4: Somewhat Agree; 5: Agree; and 6: Strongly Agree. From this kind of answer, it is possible to classify relations in a simpler manner: 0 if the person doesn't know or disagrees (it happens when the value in the previous scale is 0 or between 1 and 3), 1 if the person agrees (hence when the value in the previous scale is between 4 and 6). The graph analysis has been performed according to this latter classification of the relations.

In addition to the relational data, the dataset also contains information about the people (nodal attributes). The following attributes are known for the manufacturing company: *location* (1: Paris; 2: Frankfurt; 3: Warsaw; 4: Geneva), *tenure* (1: 1-12 months; 2: 13-36 months; 3: 37-60 months; 4: 61+ months) and the *organisational level* (1: Global Dept Manager; 2: Local Dept Manager; 3: Project Leader; 4: Researcher).

All the 77 employees have answered the survey, but the relational data reports only the no 0 answers, which stand for lack of contact between two persons, and thus absence of the edge. For the nature of the survey this is a directional network because two persons can have different opinions about each other.

Summarizing, this is a directed network, composed by 77 vertexes, the employees, and 1842 edges, which represent the number of pairs where at least one has a positive opinion of the other.

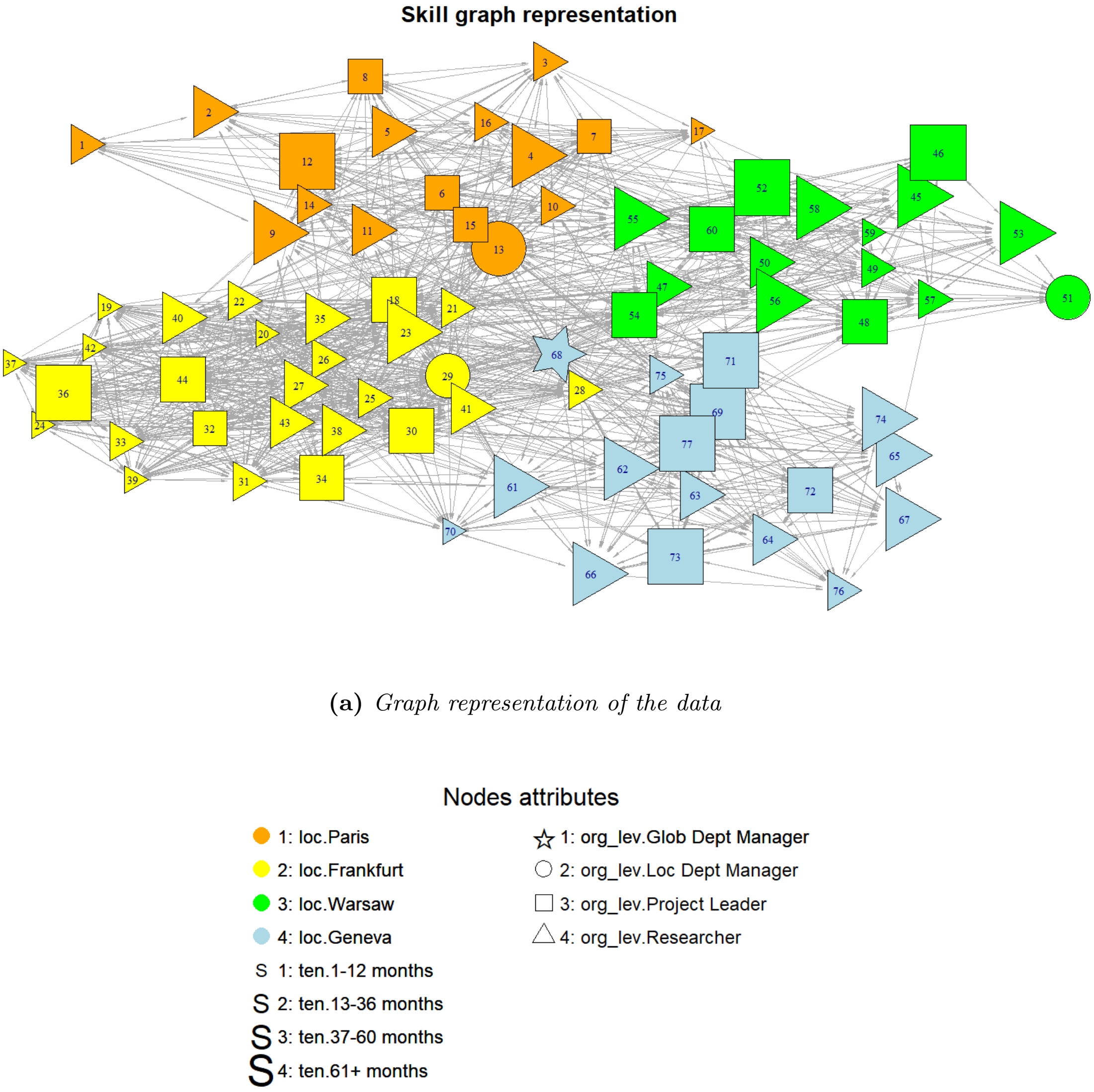
The graph (1a) appears a bit messy because of the high density (0.3148). However, it is evident the higher number of edges between employees located in Frankfurt rather than in the other three cities, above all Geneva. Moreover, it seems that nodes with small size, corresponding to new employees, are more isolated.

Nodal Features

Now we analyse the network from a Nodal Features point of view, using the main four centrality measures: Degree Centrality, Closeness Centrality, Betweenness Centrality and Eigenvector Centrality.

These measures provide different type of information about the nodes. In particular they assess how central is a node, according to different definitions of centrality. Each of this statistic can be standardized using the maximum observed value, with the

[1] R. Cross and A. Parker, "The Hidden Power of Social Networks." Harvard Business School Press (2004).



(a) *Graph representation of the data*

(b) *Graph legend*

Figure 1: *Caption for this figure with two images*

aim of comparing different networks. In addition, it is possible to obtain a Centralization Index that summarises how centralized is the network, in particular it provides information on how distant is the network to a star configuration, i.e. in which there is a single node assuming the maximum value and all the other nodes assume the same low value, or to a circle configuration, i.e. with all nodes with the same centrality.

In Table 1 we show the summaries of the standardized statistics we obtained from the network and below we provide how we got the values and some comments on them.

In and Out Degree Centrality defines central a node with many incoming or outgoing relations respectively. The measure for node i is the simple count of observed ties incoming in it or outgoing from it. The maximum observable value is when a node has incoming ties with all the other nodes in the network. The standardized statistics are, then:

$$\tilde{\zeta}_i^{in-d} = \frac{\zeta_i^{in-d}}{n-1}$$

$$\tilde{\zeta}_i^{out-d} = \frac{\zeta_i^{out-d}}{n-1}$$

The summary shows that nodes have from moderately low to moderately high centrality for both statistics. According to In Degree centrality this means that there are not workers in the company that are said to be very much more skilled than the

	In-Degree	Out-Degree	In-Closeness	Out-Closeness	Betweenness	Eigenvector
Min.	0.1316	0.0263	0.4578	0.4199	0.000007	0.1550
1st Qu.	0.2237	0.2105	0.5278	0.5241	0.0016	0.2985
Median	0.3158	0.3026	0.5547	0.5801	0.0048	0.4489
Mean	0.3148	0.3148	0.5668	0.5724	0.0104	0.4856
3rd Qu.	0.3684	0.4079	0.5984	0.6281	0.0119	0.6662
Max.	0.7105	0.7105	0.7755	0.7755	0.1305	1.0000

Table 1: *Summaries of the four standardized statistics.*

others. On the other hand, according to Out Degree Centrality this distribution underlines that people in the company rarely acknowledge colleagues skills.

In Figure **??** are shown the graphs in which the size of the vertices are proportional to the values of the In (2a) and Out(2b) Degree Centrality. The three most important nodes are coloured in red. The shapes of the nodes represent the organisation level for each node: triangles are Researchers, squares are Project Leaders, circles are Local Department Managers and stars are Global Department Managers. In both graphs we can see that the biggest nodes are mainly researchers and project leaders. This is very reasonable because they should be known in the company and maybe they have a good reputation(Fig. 2a). At the same time if they are in contact with many workers they are more aware of the skills of their colleagues (Fig. 2b). Other nodes that have a not a big size represent all nodes that are not so central according to this statistics.

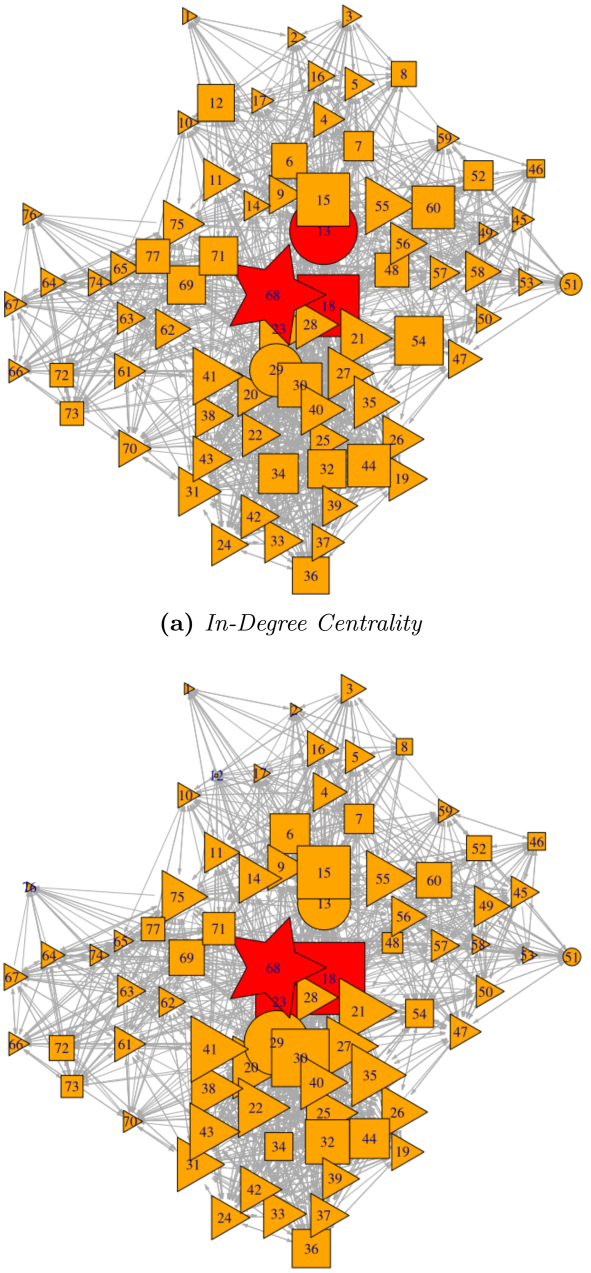
In and Out Closeness Centrality define central a node if it is close, in terms of geodesic distance, to many others. The two measures for node i are the simple sum of all geodesic distances starting from and arriving at node i. The maximum observable value is reached when the graph assume a star configuration, in which all nodes go into a central one according to the In-Closeness, or a single node goes into all the others according to the Out-Closeness. The standardized statistics are:

$$\tilde{\zeta}_i^{in-c} = (\zeta_i^{in-c})(n-1)$$

$$\tilde{\zeta}_i^{out-c} = (\zeta_i^{out-c})(n-1)$$

The two statistics provide almost the same information, and they have a very similar distribution, both assuming moderately high values.

In Figures 3a and 3b we show the graphs in which the size of the nodes are proportional to the values of the two statistics. As we can see all the nodes have quite the same dimension, there is not a large difference in centrality of the nodes for the two statistics.

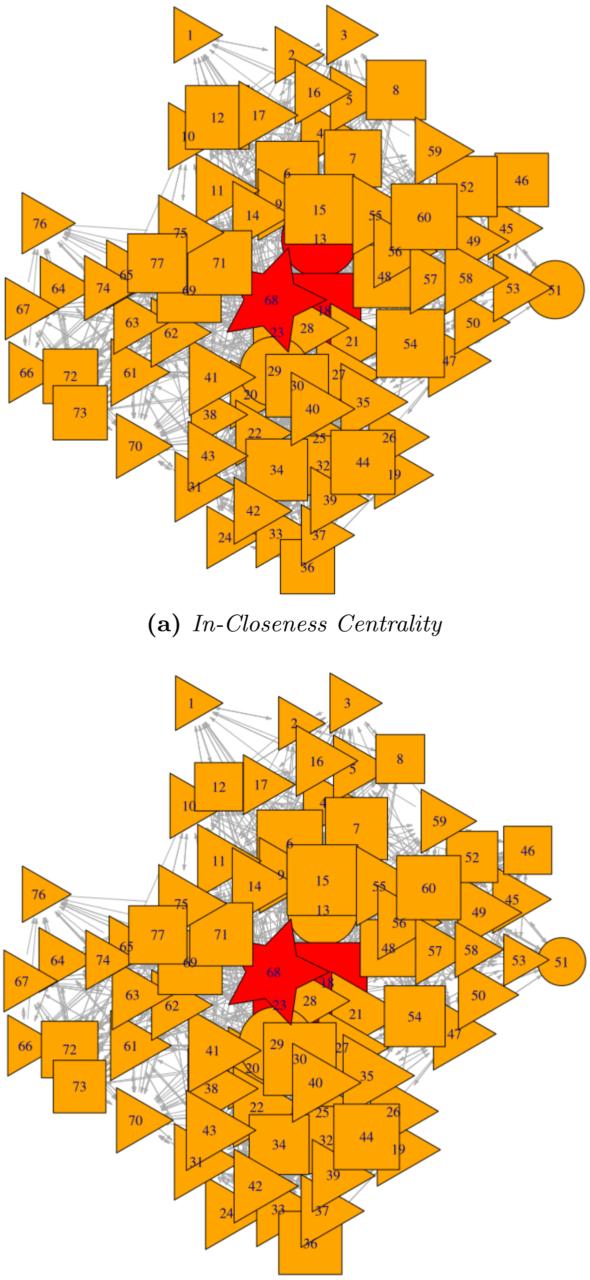


(a) *In-Degree Centrality*

(b) *Out-Degree Centrality*

Figure 2: *Nodes size proportional to the values of the Degree*

Betweenness Centrality define central a node if it is located between many other nodes. The measure



(a) *In-Closeness Centrality*

(b) *Out-Closeness Centrality*

Figure 3: *Nodes size proportional to the values of the Closeness*

is based on geodesic paths, in particular it is a ratio between the number of geodesic paths from j to k that involve node i and the number of all geodesic paths from j to k. The maximum is reached when these two values are equal, so when a node is involved in all the paths. The standardized statistic is:

$$\tilde{\zeta}_i^b = \frac{\zeta_i^b}{(n-1)^2(n-2)}$$

The statistic assume very low values, so because nodes have all a very low centrality we can say that there is not a node through which many relations pass.

In Figure 4 we show the graph in which the size of the nodes are proportional to the values of this statistic. As we can see values are so low that the nodes are very small and even the most central nodes can be barely seen.

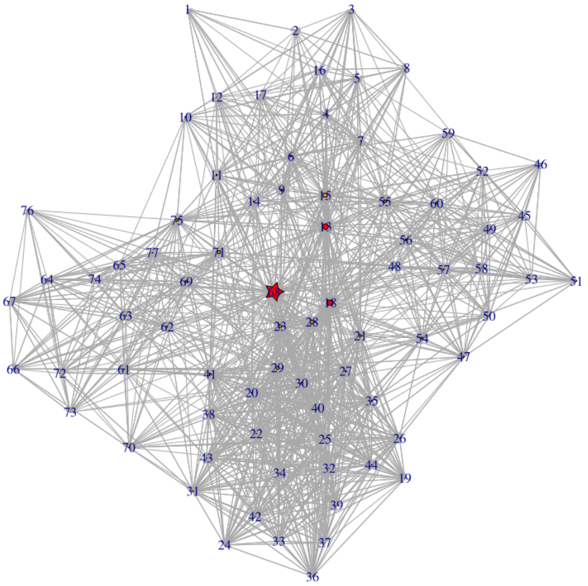


Figure 4: *Nodes size proportional to the values of Betweenness Centrality*

Eigenvector Centrality defines central a node that is connected to other central nodes. As the name suggests, the measure involves the calculus of eigenvectors and eigenvalues.

The distribution of the statistic has a wide range, nodes have from moderately low to high centrality. Most nodes have a moderately high centrality, but there is a node who reaches the maximum value for the statistic.

As we can see in Figure 5 some nodes have a low centrality, which are represented on the boundaries of the graph, but there is a group of nodes with higher centrality. This group is the same underlined during the analysis of the Degree Centrality but the differences between the group and the others are bigger. It is composed by Researchers and Project Leaders, that reasonably have more contacts with important people in the company.

Most central nodes The most popular nodes in the network shown in Table 2 are the nodes that result the most central according to all statistics. The four nodes belongs to different branches of the company and they are all 'old' employers and they cover important roles in the company.

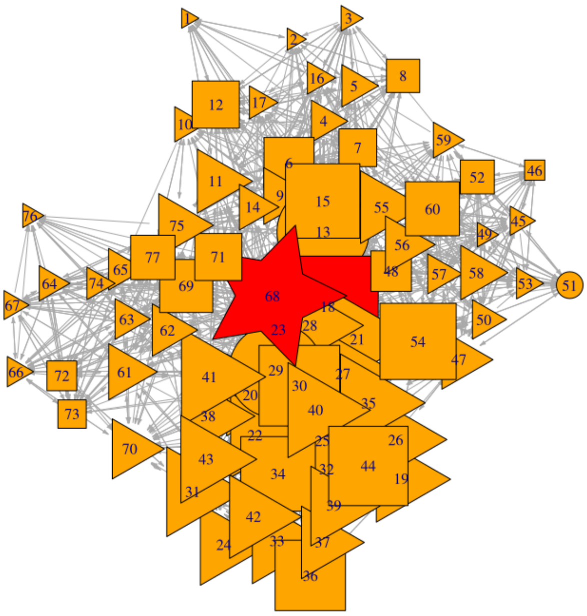


Figure 5: *Nodes size proportional to the values of Eigenvector Centrality*

Node	Location	Tenure	Level
68	Geneva	37-60 months	Glo. Manager
13	Paris	61+ months	Loc. Manager
18	Frankfurt	37-60 months	Proj. Leader
23	Frankfurt	61+ months	Researcher

Table 2: *Most central nodes*

Given all these features, it is reasonable that they are the most central nodes in the network. In addition they could be the bosses of the company.

Centralization Indices are reported in Table 3 providing the formulas and the values obtained.

The values for In- and Out-Degree Centralization Indices are the same, in particular they both assume 0.401. This value means that the network is not close neither to a circle configuration, the least centralized, neither to a star configuration, the most centralized. The network seem to be an half way between this two configuration.

Even the values for In- and Out-Closeness Centralization Indices are very similar, but they are lower than the values obtained for the Degree Centrality. More precisely they assume values near to 0.21. This values means that the network is less centralized and so nearer to the circle configuration according to this definition than before. There is not a node that has a much higher centrality than the others, instead all nodes have almost the same values for centrality.

For the Betweenness Centralization Index we observe an even lower value than the others. As we have already said, according to this definition of centrality nodes were more similar to each other all assuming very low values. This result shows that the network is not centralized at all, the value is very near to the one that we would obtain if we had a circle configuration.

Index	Formula	Value
CI^{in-d}	$\frac{\sum_i \zeta_{max}^{in-d} - \zeta_i^{in-d})}{n-1}$	0.401
CI^{out-d}	$\frac{\sum_i \zeta_{max}^{out-d} - \zeta_i^{out-d})}{n-1}$	0.401
CI^{in-c}	$\left[\sum_i \zeta_{max}^{in-d} - \zeta_i^{in-d}\right](n-1)$	0.214
CI^{out-c}	$\left[\sum_i \zeta_{max}^{in-d} - \zeta_i^{in-d}\right](n-1)$	0.208
CI^{b}	$\frac{\sum_i \zeta_{max}^{in-d} - \zeta_i^{in-d}}{(n-1)^2(n-2)}$	0.122

Table 3: *Centralization Indices for the statistics.*