Complex Networks

Marco Murtinu, Claudio Peroni

June 16, 2018

Introduction

The dataset we are going to analyze contains information about the e-mail communications between 1005 people working in the same university. For each person we know to which department of the university they belong. There are 42 departments: we do not know the departments' names, they are just labeled with the integers in the range $[0, 41]$.

Since each node (person) can send as well as receive mails, the natural way to model this situation is a directed graph where each node respresents a person and each edge represents an email beetwen two people: particularly, for a node n, an inner-edge represents the receipt of an email, while an outer-edge means that he sent an email. Each node possesses the attribute *department* which represents, of course, their department (as an integer).

It is good practice to start any data analysis by cleaning the data: in this case we decided to remove all the self-loops in the graph. Indeed, they indicate the mails to one self, which are irrelevant to the analysis[1]. The resulting graph, is visible in Figure 1.

We also created an alternative, undirected version of the graph: we will see that this will be useful for some analysis and some interesting comparisons in the proceeding of this work. Clearly, we are aware that this two versions of the graph do have some differences, the most notable one being the numbers of edges and, consequently, the degree of nodes.

DIRECTED VERSION:

- Number of nodes: 1005
- Number of edges: 24929
- Average in degree: 24.8050
- Average out degree: 24.8050

UNDIRECTED VERSION:

- Number of nodes: 1005
- Number of edges: 16064
- Average degree: 31.9682

[1]Supposing the analysis is centered around actual relationships between people and departments, and not something like mail utilization.

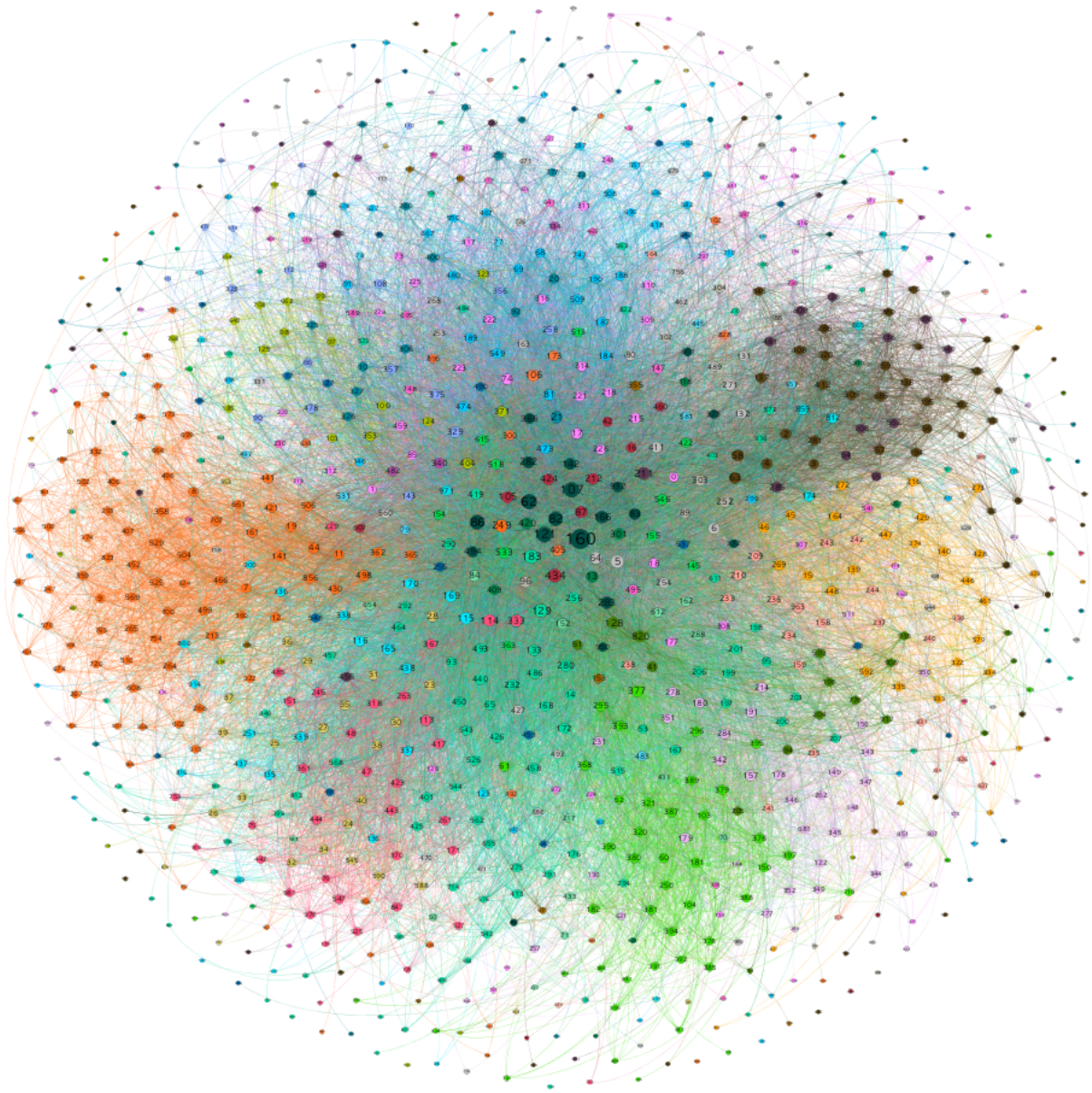

Figure 1: Emails graph: the size of nodes is proportional to their degree, the colors represent the modularity class assigned by Gephi. The colors of the edges denote the modularity class of the source.

We created another related graph, visible in Figure 2, representing only the departments: this is an undirected weighted graph having the departments as nodes, where the edge between two departments has a weight proportional to the number of communications between their elements. We will see that this graph will be useful to draw further conclusions, looking at the data with an aggregate perspective. Here is a summary of this graph:

- Number of nodes: 42
- Number of edges: 679
- Average degree: 32.3333

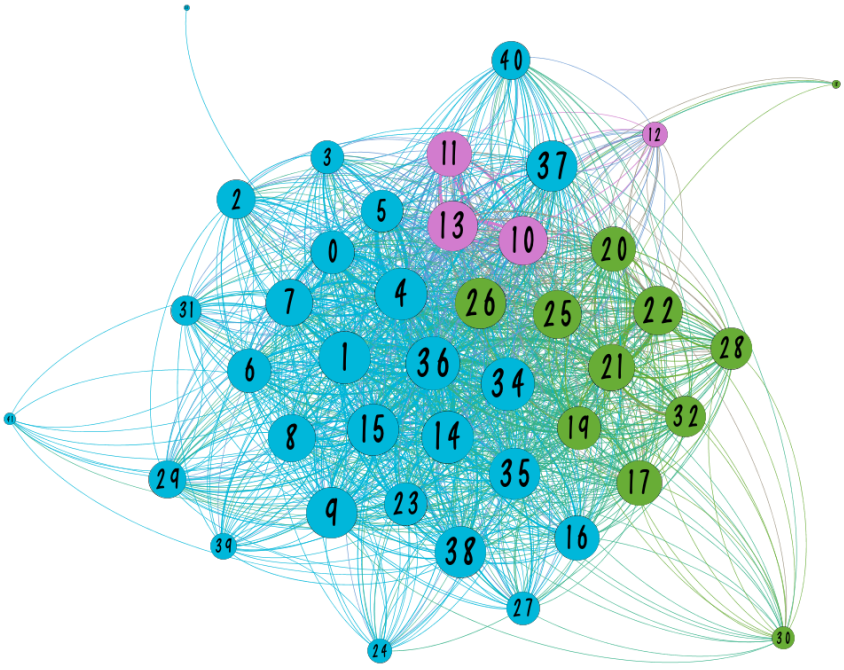

Figure 2: Departments graph: the size of nodes is proportional to their weighted degree, the colors represent the modularity class assigned by Gephi.

Connected components and Giant component

We begin our study by analyzing the structure of the network.

First, we search for the weakly and strongly connected components in both the directed and the undirected mail graphs.

- Number of connected components (undirected version): 20
- Number of weakly connected components: 20
- Number of strongly connected components: 203

Clearly, the number of connected components of the undirected graph is the same as the number of weakly connected components of the directed graph, as expected by definition.

It comes as no surprise that the strongly connected components are much more than the weakly connected components, since the latter is a easier requirements to meet: recall that a directed graph is said to be **strongly connected** if any vertex is reacheable from every other vertex, whereas a directed graph is **weakly connected** if, by substituting all its edges with undirected edges, we obtain a connected graph.

Speaking about connected components, it is of considerable interest to understand whether there exists a **giant component**, i.e., a connected component that contains a significant fraction of all nodes. Let us start from the undirected version of the graph, which has only 20 connected components.

COMPONENT 0:

- Number of nodes: 986
- Number of edges: 16064
- Relative Order (proportion of nodes): 0.9811
- Relative Size (proportion of edges): 1.0

COMPONENT 1:

- Number of nodes: 1
- Number of edges: 0
- Relative Order (proportion of nodes): 0.0010
- Relative Size (proportion of edges): 0.0

COMPONENT 2:

- Number of nodes: 1

- Number of edges: 0
- Relative Order (proportion of nodes): 0.0010
- Relative Size (proportion of edges): 0.0

⋮

COMPONENT 19:

- Number of nodes: 1
- Number of edges: 0
- Relative Order (proportion of nodes): 0.0010
- Relative Size (proportion of edges): 0.0

It is easy to see that the first component represents a giant component: indeed it contains more than 98% of all the nodes and all the edges of the graph, whereas the other 19 components are all composed by a single node (they are what we may call detached leaves).

Let us now try to find a giant component also in the directed graph.

COMPONENT 0:

- Number of nodes: 1
- Number of edges: 0
- Relative Order (proportion of nodes): 0.0010
- Relative Size (proportion of edges): 0.0

COMPONENT 1:

- Number of nodes: 1
- Number of edges: 0
- Relative Order (proportion of nodes): 0.0010
- Relative Size (proportion of edges): 0.0

⋮

COMPONENT 162:

- Number of nodes: 803

- Number of edges: 24138
- Relative Order (proportion of nodes): 0.7990
- Relative Size (proportion of edges): 0.9683

⋮

COMPONENT 202:

- Number of nodes: 1
- Number of edges: 0
- Relative Order (proportion of nodes): 0.0010
- Relative Size (proportion of edges): 0.0

Even in this case we easily recognize the giant component to be the component #162: it contains around 80% of the nodes and more than the 96% of all the edges of the graph. The other 201 components are composed just by a single node. This is the best situation we could hope for, in a sense, since it enable us to do further analysis that require connectedness without having to discard a significant portion of the network.

Eventually, we calculated also the number of connected components in the departments network and, as expected by the nature of it, there is only one such component, which means that it is possible to find a communication path from each department to any other (thakfully).

Degree

Let us now study the distribution of the degree among nodes.

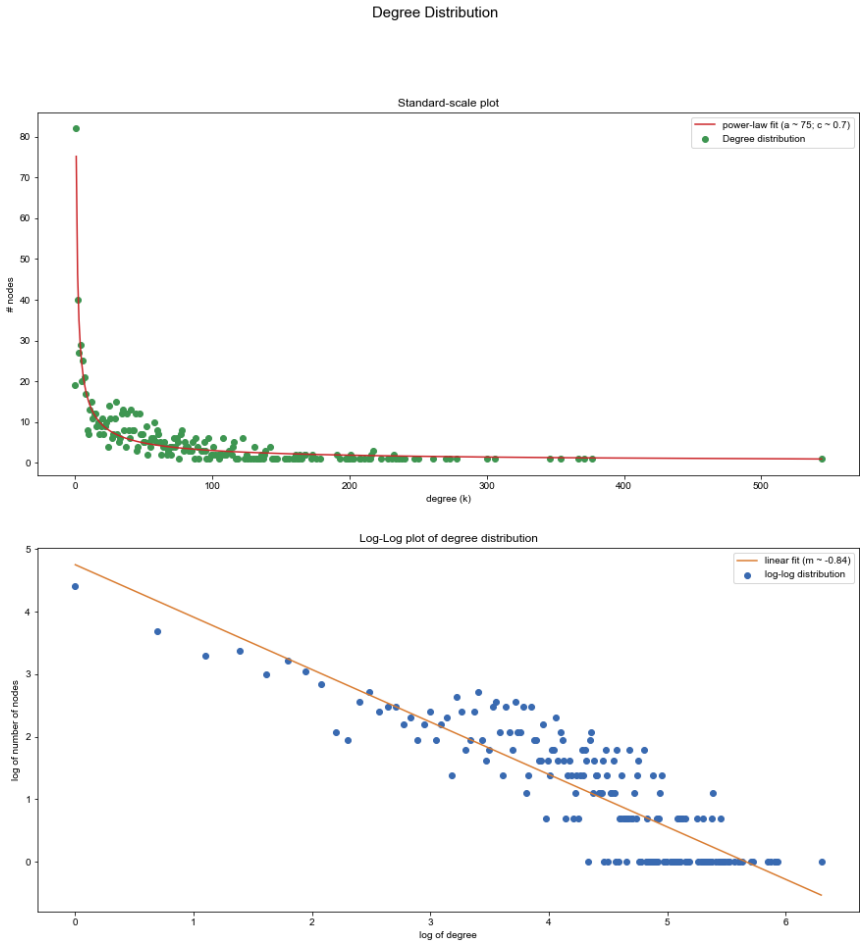

Figure 3: The first graph represents the degree distribution: its shape resembles a power law, therefore, in order to confirm it, we plot it again after a log transformation. The red lines represent the theoretical power-law that best fits the data. In both cases we can say that the data roughly follow a power-law, at least to some extent.

The first plot in Figure 3 shows the degree distribution of the nodes: its shape roughly resembles a power-law, which may be reasonable since the degree of a node can be intended as a kind of popularity measure of the node itself. However the first plot is not enough to fully confirm such hypothesis, therefore

we apply a logarithmic transformation. Indeed, if $f(k)$, the function representing the frequency of the degree k among nodes, follows a power-law, then

$$f(k) = \frac{a}{k^c}$$

where a is a constant of proportionality. Our aim is to find c. Hence we apply a log transformation, so to have

$$\log(f(k)) = \log(a) - c\log(k)$$

In this setting, a is the intercept with the y axis and c is the slope of the line. In the second plot we can roughly see a line with slope between 1 and 5/6, which supports the power-law hypothesis to some extent.

We also compute the average degree, the maximum degree and the standard deviation of this distribution

- Average degree : 49.6099
- Standard deviation: 59.7496
- Maximum degree: 544[2]

It may be of interest to study the degree's distribution also in the department graph. Note that, differently from what we have found in the emails graph, the plots in Figure 4 strongly deviate from the theoretical behavior of a power-law.

[2]Note that this degree belongs to node #160, which will appear again and again in this analysis, being the most central node in the graph with a degree way above that of the others.

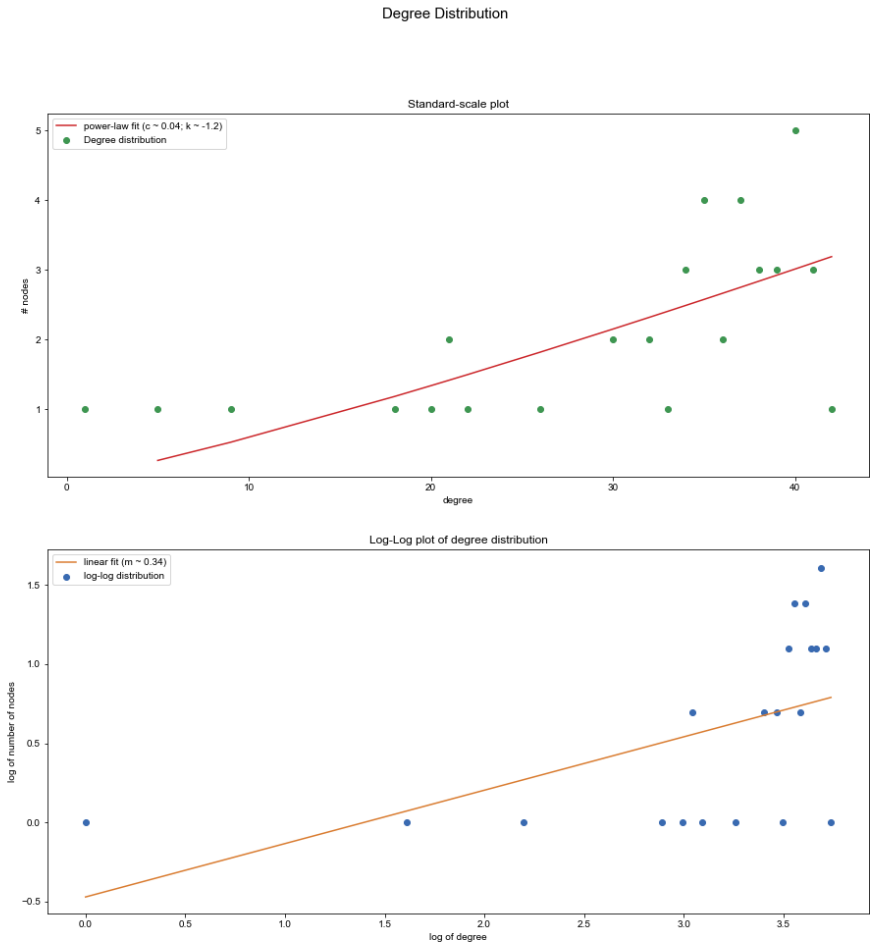

Figure 4: Degree distribution in the departments graph: in the first plot we have the degrees on the x-axis and their frequences on the y-axis. In the second one we applied a log transformation to better evaluate the presence of a power law distribution, which is clearly not present.

Distances

We now compute the distribution of distances: it is important to stress that in this analysis we are bound to consider only the *giant strongly connected component* for the directed graph and the *giant connected component* for the undirected graph, otherwise we would obtain infinite values. Let us first study the directed graph.

The distribution of distances in the *giant strongly connected component* is shown in Figure 5.

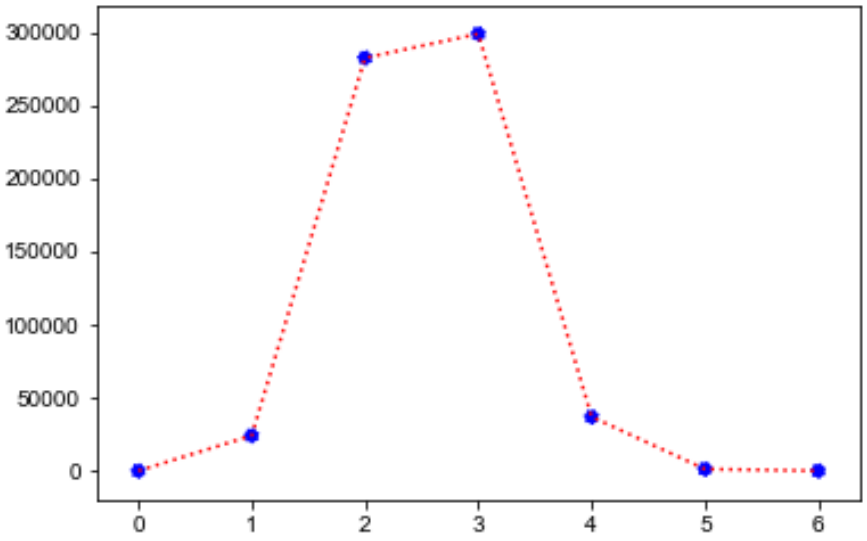

Figure 5: Distance distribution in the directed graph.

We then compute some statistics of such distribution[3]:

- Average distance in the strongly connected component: 2.5443
- Standard deviation of distance in the strongly connected component: 0.6757

Let us now do the same for the undirected graph, as seen in Figure 6.

[3]Note that for this analysis we decided to exclude the null distances, *i.e.* all the distances between any node and itself, which will always be the same number as the number of nodes.

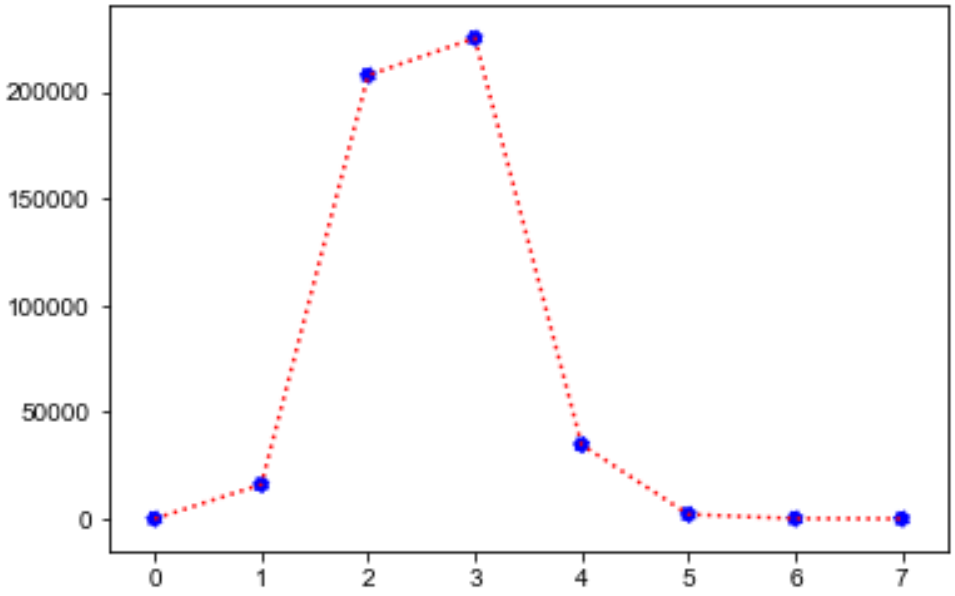

Figure 6: Distance distribution in the undirected graph.

Once more, we compute the average and standard deviation of the distribution:

- Average distance in the connected component: 2.5843
- Standard deviation of distance in the connected component: 0.6970

To conclude our discussion about distances in the emails network, we compute the *eccentricities* in the strongly giant component as well as in the giant component. In both cases, we report the diameter and the radius of the networks (which are the *max* and *min* eccentricities respectively).

- Radius of the giant strongly connected component: 3
- Radius of the giant connected component: 4
- Diameter of the giant strongly connected component: 6
- Diameter of the giant connected component: 7

We see that the giant strongly connected component (with $\sim 80\%$ of the nodes of the whole graph) effectively has a smaller radius and diameter than the giant connected component of the undirected graph (which has $\sim 98\%$ of the nodes).

Once more, we conclude the analysis by studying the distance distribution also on the departments graph.

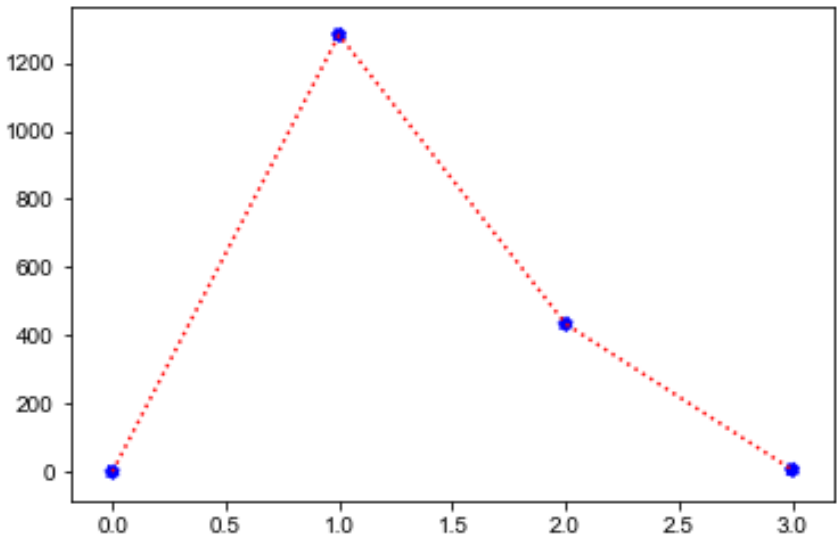

Figure 7: Distance distribution in the departments graph.

It is apparent how different the distribution of the distances in the *mail network* and the one in the *department network* are. This comes to no surprise, considering that for two departments to be neighbors it is sufficient that any two people inside them exchanged a single mail.

Centrality

Understand which groups or single nodes are central for a network is not immediate: indeed there are several different measures of centrality, each focusing on a slightly different aspect and proposing and slightly different notion of centrality. However we will see that in our graph all these measures will agree in identifying the central group and the central node.

The first measure we compute is probably the simplest and most intuitive one, the **degree centrality**, which measures the centrality of a node by its degree, i.e., by the number of nodes connected to it. Unsurprisingly, coherently with that we obtained when we studied the degree distribution, we see that node #160 is the most central node. Moreover, node #160 as well as most of the other most central nodes, belongs to department 36, which is identified as the most central department.

A second measure of centrality is the **eigenvector centrality**: the idea underneath this measure is that a high eigenvector score means that a node is connected to many nodes who have high scores themselves. Despite using a

different measure, we still find that node #160, belonging to department 36, is the one with highest eigenvector centrality.

Another possible measure is the **closeness centrality**, calculated as the sum of the lengths of the shortest paths between the node and all other nodes in the graph. For a node n

$$C(n) = \sum_{m \neq n} \frac{1}{d(n,m)}$$

where m is another node in the graph, is its closeness centrality. Thus the more central a node is, the closer it is to all other nodes. Once more, node #160 is the one with the highest centrality measure.

Eventually, we compute the **betweeness centrality**, defined as the number of shortest paths that pass through a node, i.e.,

$$b(n) = \sum_{m \neq n \neq l} \frac{\sigma_{m,l}(n)}{\sigma_{m,l}}$$

where $\sigma_{m,l}$ is the total number of shortest paths from node m to node l and $\sigma_{m,l}(n)$ is the number of those paths that pass through n. Also in this case, node #160 receives the highest score.

However, the situation becomes less clear when analyzing the centrality of the departments network: we are going to see that the centrality measures on this graph will support our previous conclusions only to some extent.

We have a first confirmation of the centrality of the department 36, that we discovered in the emails network, by using the *degree centrality*: this department scores a value slightly higher than departments 1 and 15. The situation becomes less clear when computing the betwness centrality: in this case the department 4 gets a value significantly higher (0.0551) than the value of department 36 (0.0223).

Clustering coefficient

First, we compute the average clustering coefficient for the whole graph. Note that in order to do this we need to use the undirected version.

- Average clustering for undirected graph: 0.3994

We are also interested in checking the clustering coefficient of the most central node, #160, which is 0.0935.

Note that the fact that the clustering coefficient of the most central node is very low makes sense: since it is connected to huge number of nodes, it would be strange if a lot of them were connected to each other; indeed, they can be far away from each other.

Let us now exploit the fact that we are aware of a natural partition of the nodes in the graph: departments. It may be interesting to see whether such

natural groups form strongly clustered regions. Recalling that the average clustering coefficient of the network is around 0.4, we see that most departments, with the expection of those composed by just one node, which of course score 0, and those composed by just 2 nodes with a link between them, which clearly score 1, have a clustering coefficient significantly above the average.

Density

Another measure of the thickness of the graph is its **density**, defined as the ratio between the number of edges between all the nodes and the number of all the possible edges between such nodes. We computed it for both the emails and the departments networks.

- Density of email network: 0.0247
- Density of department network: 0.7886

As we might expect, the departments network is much more dense than the emails graph.

Homophily

We would like to understand whether our network exhibits some sort of *homophily*. However, this analysis is difficult in our situation, becausee we do not know anything about the single node: the only information that we have about a person is his or her department, we do not know anything about his age, gender, nationality or other personal features with respect to analyze the homophily. Therefore, we are bound to try to compute the homophily with respect to the only information we have, the department. Note that this a "formal" homophily: we will basically understand whether people in the same department communicate more often with people within their department than with people of other departments.

To analyze homophily we consider a network in which the attribute *department* is randomly assigned to each node with probability proportional to the proportion of nodes belonging to that department in the original network. Therefore, denoting by p_i the probability that a node belongs to department i, and considering two random nodes linked by an edge, the probability that they belong to the same department is

$$\sum_{i=0}^{41} p_i^2$$

Hence the probability that they belong to different departments is

$$1 - \sum_{i=0}^{41} p_i^2 = 0.9524$$

Instead, considering our network and counting the proportion of edges between nodes belonging to different departments, we obtain 0.6643. This shows that in our graph there are significantly less edges that connect nodes from different departments than in a random graph with the same proportion of nodes in each department. Hence we can say that our network displays homophily.

Assortativity

The **assortativity** represents the preference of a network's nodes to attach to others that are similar in some way. We operationalize this idea by measuring assortativity, as it is common practice, with respect to nodes' degree.

The plot in Figure 8 clearly shows that there is no evidence for assortativity: indeed, in such a case, we would expect an increasing trend of the curve. By the same token, in case of dissortativity, we would expect a decreasing trend in the curve. In this case, we cannot see any of these behaviors, hence we must conclude that our network is mostly neutral with respect to assortativity[4].

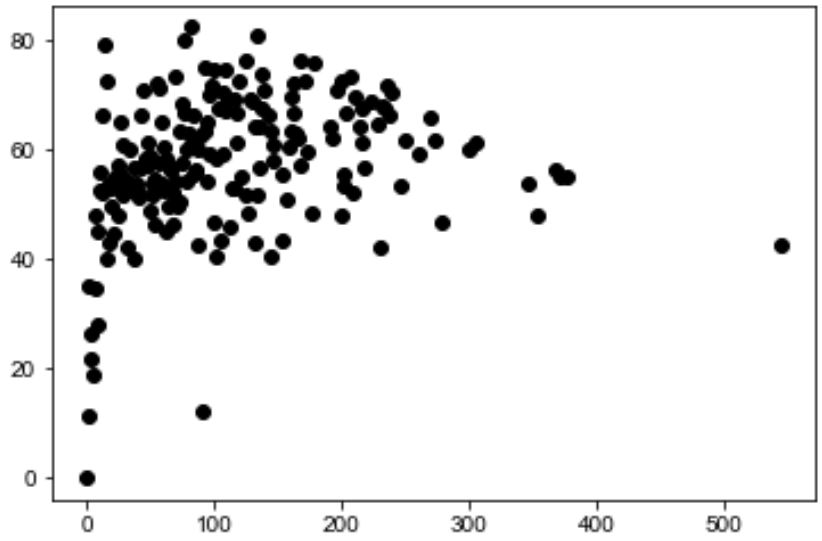

Figure 8: Assortativity in the emails graph: there is no clear evidence neither for assortativity nor for dissortativity. We may conclude that the network is neutral.

As we have done for all the other analysis, let us see whether we can find assortativity in the departments graph. We can see that, differently from the

[4]Actually one could argue that the network is assortative only for very small degrees, but due to the lack of more data in this regard, this assertion remains in the realm of speculation.

emails network, the graph shows some dissortativity, having a degree assortativity coefficient of -0.1786. This suggests that departments with high degree are slightly incline to attach to departments with lower degree.

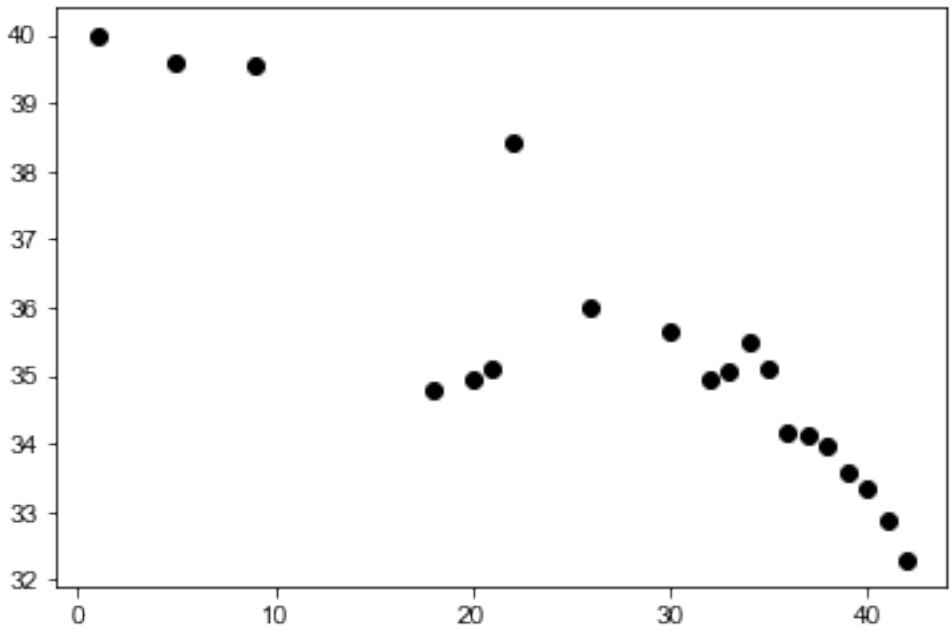

Figure 9: Assortativity in the departments graph. The plot shows a certain degree of disassortativity.

Conclusions

Let us now draw some conclusion about our network. In fact, we have discovered three interesting features of both mails and departments graph: first of all, we have seen that the degree distribution of the emails graph roughly follows a power-law, which is, in some way, a comforting news, since we know that such law is pervasive when studying notions of popularity.

We have also been able to identify clearly a central node and a central department, number 160 and number 36, respectively. Although there are no data to confirm it, we could speculate that the department 36 has an administrative role and that node 160 occupies a key position within it.

The third interesting feature we have found out concerns the homophily of the mails network: as expected, people tend to communicate more often with other people within the same department. Unfortunately we cannot proceed further with only these information, however if we would be able to obtain at least the general research area of each department (like 'scientifical' or 'medical' or 'humanistic'), it would be interesting to see whether people actually display

homophily also with respect to the research area (it may be interesting to verify whether, mathematicians tend to be more in contact with physicists than with paleontologists, for instance).