Final report: Challenge 2

Final report of the Graph Theory course

Group : Daniela Batalha and Ishita Chaudhary
Teachers : Cécile Bothorel ; Yannis Haralambous ; Nicolas Jullien

IMT Nord Europe
École Mines-Télécom
IMT-Université de Lille

Contents

Evaluation of Organization's performance based on Email Interactions

1. INTRODUCTION

In this challenge, we try to examine the organizational structure of a company by evaluating the internal communication (based on email interactions) at an organizational level by applying cluster techniques and network analyses. In our study, we are concerned with internal communication in an organization. The purpose of our analyses is to examine how email affects work performance. For that, we have done treatments in the data set and have also used cluster and network analyses to identify the possible communities into the organization. In addition, we calculated the highest measures of centrality into each community to identity patterns into the network and propose strategies to improve performance of the organization, based on email analysis.

2. PROBLEM EXPLICITATION

a. E-mail and work performance

Email is an increasingly important tool of communication that facilitates contact between individuals and enable to rise in productivity and optimization of processes in companies. In the present world, a huge volume of data is being generated from the exchange of email between people in companies, and it has enabled organizations to have more spread-out and diverse work teams. In addition, the use of email in organizations can be useful to improve management process by allowing interdepartmental communication [1] . For example, in global companies located in many countries, email allows employees to communicate with each other across country borders.

However, even if there are many positive points related to the power of email exchange in companies, the misusing of this tool can result in different problems into organizations. We can mention, for example, lower productivity, because of the almost uncontrollable flow resulting in information overload [1]. A study realized in the U.K. showed that that corporate e-mail users spend at least five minutes dealing with each message; so that, in a work day, more than three hours, on average, are spent dealing with e-mail [1]. The impact of it is a frequent interruption in the work routine, affecting the normal work flow of the employee.

Another problem is that emails naturally increase the number of tasks that users are expected to cope with [1]. One positive consequence of is the possibility to make important network connections with other workers. However, the negative effect is the excessive assignment of

tasks related or not to the employee's role. Therefore, these excessive attribution of tasks, exposed through the volume of email that a worker can receive, can strongly cause effects on the user's well-being and performance [1].
One other problem related to emails shared in companies is the cost of data storage. Electronic storing space can become a problem, particularly where emails with large attachments are widely distributed. To help companies to reduce the email size impact, virtual work teams tools can be used to share of archives, for example. Also, it's useful when it's necessary a fast communication.
Considering the effects of email communication detailed above, Human Resources managers can invoke email communication analyses to understand the behavior of employees, with objective to improve individual and organizational performance. For that, social network analyses can be implemented, by using graph theory concepts to identity the different aspects of the relationship existent among employees. By understanding the communication process, companies can optimize networks communication that rise organization's productivity.

In this challenge, we are interesting to use social network analyses to understand the internal communication among employees, examining how email exchanges can affect work performance. For that we are interested to see workers that send/receive a lot of emails, time that emails are being send, size of emails, the impact of the hierarchy in email communication, groups of employees, among others. All these analyses will be useful to understand the workers' company behavior and propose some tolls to improve performance into the company

b. Corporate Social network analyses

Social networks methods are often used to examine how organization employees interact among them. Social network analysis produces an alternative view that can provide additional knowledge about the flow of communication into the company, that can be further used to improve the management of the company in various ways [1]. Also, network analyses can be useful to identity key people inside a company, to verify it employees are overloaded of job or, the opposite, if the worker has not been productive, for example.

For this analysis, we reformulate the problem as a directed weighted graph G := (V, E), where we have the organization's employees (in the node-set V).
Nodes represent the employees and edges represent the message shared between them. The edge between any two nodes will be directed from the sender towards the receiver. Each edge will have an ID associated with it; the subject and date-time stamp will determine the ID to identify if the same email was sent to many recipients. The weight of edge e_{ij} , from node i to node j is defined as:

$$w_{ij} = \frac{number\ of\ emails\ sent\ from\ \ worker\ i\ to\ j}{total\ number\ of\ emails\ sent\ from\ worker\ i}$$

We will use a different color of nodes for different seniority levels of nodes, for observation purposes. Also, we will calculate a local clustering coefficient for each node, which allows capturing the density of connections between neighbors, as well as two additional features related to cliques:

$$CCi = 2(\text{number of pairs of neighbors of a node i})/(d(d-1))$$

where d is the degree of node i.

3. DATA PROCESSING

a. Data Understanding

The dataset used in this analysis comes from a consulting firm and the content represent emails exchanged during the period 4 to 19 March 2019. The total of lines is 1.174.928, and each line describes what a collaborator sent or received in a specific date ("MessDate") . When a collaborator sends an email, "Id_Direction" is equal to 1. On the other hand, when an email is received from a contact, "Id_Direction" is equal to 2. The interlocutor can be "interne"/ "externe"/ "unidentified" ("PartnerTypeName").
When "Id_Direction" is equal to 1, "PartnerName" represent the sender and "Recipient_Display" represents the receiver. On the other hand, when "Id_Direction" is equal to 2, "Recipient_Display" represents the sender and "PartnerName" represent the receiver.

The interaction involving the collaborator is defined by:

- GroupName1 : Post/Title of the collaborator
- RecipientName : The collaborator
- MessSize : message size in Mégabytes
- MessDate : Date+Hour
- Id_Direction : 1 for a sent email, 2 for a received email
- PartnerTypeName : the interlocutor may be either:
    - Interne
    - Internet (external interlocutor)
    - Unidentified local address (applications or server or cloud mail address)
- PartnerName : domain name if external interlocutor
- Recipient_Display : the interlocutor

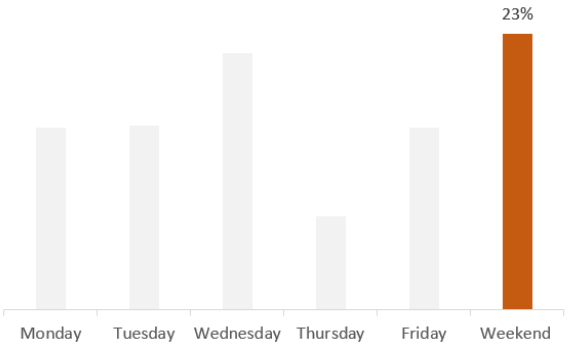
This section presents important background information involved in this study of the data set. Analyzing the data, we identify the frequency of persons per role. In table-1 we can see that the majority of employers are staff (40% of the total), followed by seniors (20 % of the total).

role	quantity
Staff	646
Senior	316
Manager	200
Junior Staff	116
Senior Manager	114
Director	65
Partner	64
Title:[no value]	51
Administrative	30
Assistant Manager	4
total	1606

Table-1

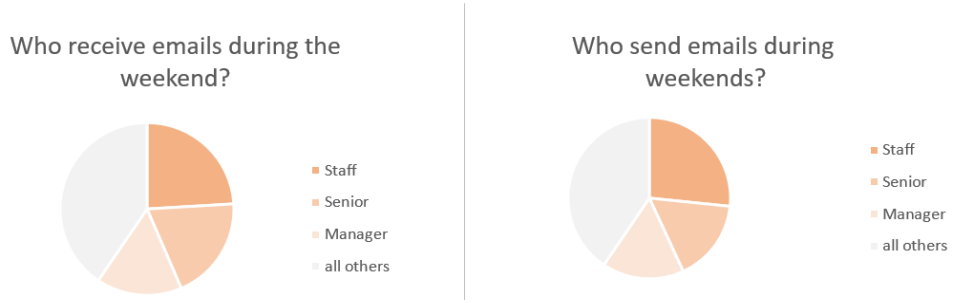
The purpose of this study is to investigate the internal communication inside this company. For that , we have selected just messages shared internally by filtering only PartnerTypeName equal to "Internal". After selecting just internal communication, the data set is reduced to 702.932 lines.

The analysis of communication by email between employees can also be useful to identity some patterns into the company. As it is demonstrated in Graph-1, 23% of internal messages communication happen during the weekend. This highlight some potential problems at Human Resource level . Are some employees been overload of job during the weekend? Is it correct to send/to receive messages during the weekend?



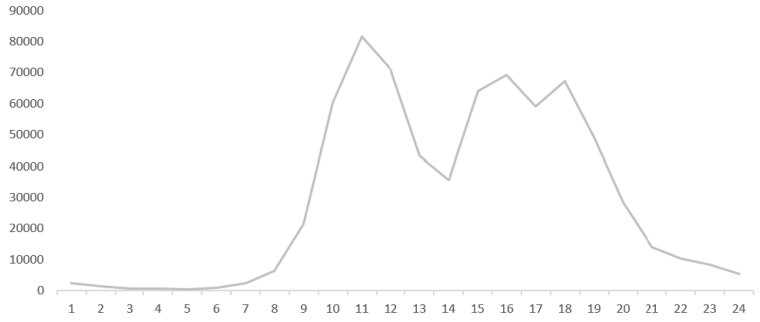
Graph-1

When we look into details about messages shared during the weekend, we can see that the roles that more send/receive messages during the weekend are staffs, seniors and managers. They represent more than 50% of messages shared during the weekend. We can see that below, in Graph-2.



Graph-2

Another useful descriptive analyses is related to the flux of messages during the day. In Graph-3 we can see that the majority of communication by email happen during work hours. The highest flux of messages is between 10am and 12am and in the middle of afternoon.



Graph -3

b. Data processing

To be able to handle with the data set, we have done some data processing.
As discussed above, we are only focusing on the emails exchanged within the company, because the emails received by the employees outside the firm cannot be controlled. As data scientists, we can only suggest how they can reduce the mails exchanged within the organization for better productivity results.

Out of total 1.174.928 lines of data, only 702.932 are "Internal" emails. Hence, we will be focusing on that. On the basis of information provided, we still need to figure out the role of sender and the role of receiver for each line for better understanding and analysis of the data.

As we already know, in case of Id_Direction=2, the GroupName is the role of collaborator (which receives the email) and Recipient_Name is also the collaborator. It is safe to say that Id_Recipient also corresponds to the ID of the receiver. So the basic idea is to build a one-on-one correspondence between Recipient_Name and ID_Recipient and GroupName. Using the Recipient_Name and ID_Recipient correspondence, we can map the Recipient_Display to the respective IDs and further the GroupName, which will give us the role of both the sender and receiver for each email.

In the code, the following dictionaries represent the key-value pair between the following datasets:

- dict_group_recID={ID: Group Name}

[*Note: The ID 1008154 corresponded to two different roles, Junior Staff and Staff, so we considered it to be a Junior Staff only*]

- dict={Recipient Name: ID}

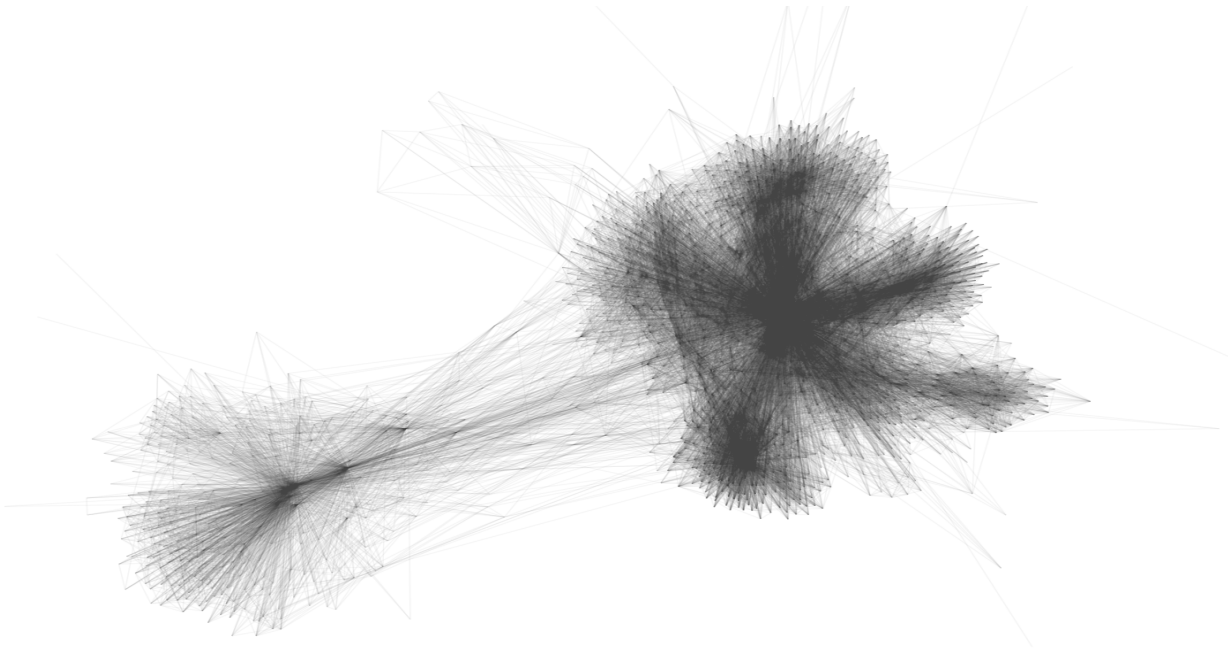
[*Note: There were a lot of entries with bijection in this case, out of 25913 unique recipient names, 17023 of them corresponded to more than one IDs. We took the most occurring ID for each recipient name to make a one-to-one mapping*]

Using these two dictionaries, we found the corresponding IDs and roles of the senders (Recipient_Display, ID_Direction =2). Out of 702932 Internal mails, 177319 had an empty mapping for Recipient Display's ID and role. But, we still include these in our analyses. Out of 702.932 mails, 389.179 are the number of emails received.

The average number of emails received per employee is
:= total internal mails received/number of employees
= 389179 /1606
= 242.32

Our goal is to reduce the average number of mails received per employee to increase productivity into the company.

The network graph below is the representation of received emails between employees.



Social network

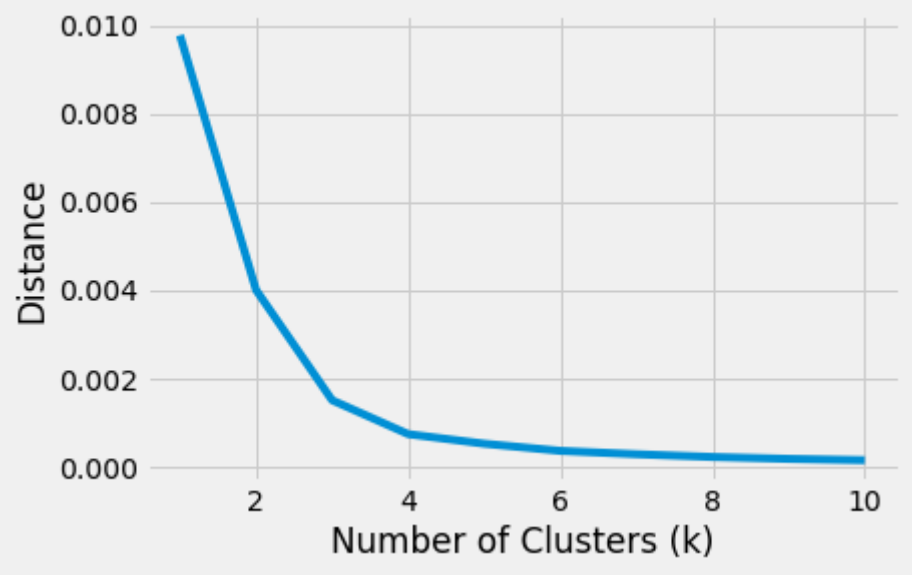
4. EXPERIMENTS

In this section, we have tried some experiments to identity the groups into the company based on email exchange between employees. Our objective is to see in details which works have more communication, which are the essential in operations, etc. The idea is to identity some patterns in the data that can be used for the Human Resources department to improve process

and performance into the company. For that, we used K-means to predict the number of possible clusters, communities detection techniques to predict and visualize the network. Finally, we calculated the centrality measure into communities to identity the most 'important' employees.

K- Means Cluster

We decided to use K-means to identity the possible quantities of clusters into the network. To process the learning data, the K-means algorithm starts with a first group of randomly selected centroids, which are used as the beginning points for every cluster, and then performs iterative (repetitive) calculations to optimize the positions of the centroids [2] . Calculating the Silhouette Index curve, we find that the optimal quantity of clusters is around 6. The Silhouette Index is a cluster validity index that is used to judge the quality of any clustering solution. We can see the Silhouette Index result below.



Silhouette Index

To go deeper in details about these groups and the patterns of these cluster, we decided to use communities detection network techniques to be able to identity the communities and the employees that are inside them.

a. Community detection with modularity

A community is typically a connected subnetwork. Qualitatively, a community is defined as a subset of nodes within the graph such that connections between the nodes are denser than connections with the rest of the network [6].
After to reformulate the problem into a graph, we decided to go deeper into the network dataset to understand what are the subgroups or communities within the larger company structure. For that, we choose the method of greedy modularity communities, that tries to determine the number of communities appropriate for the graph, and groups all nodes into

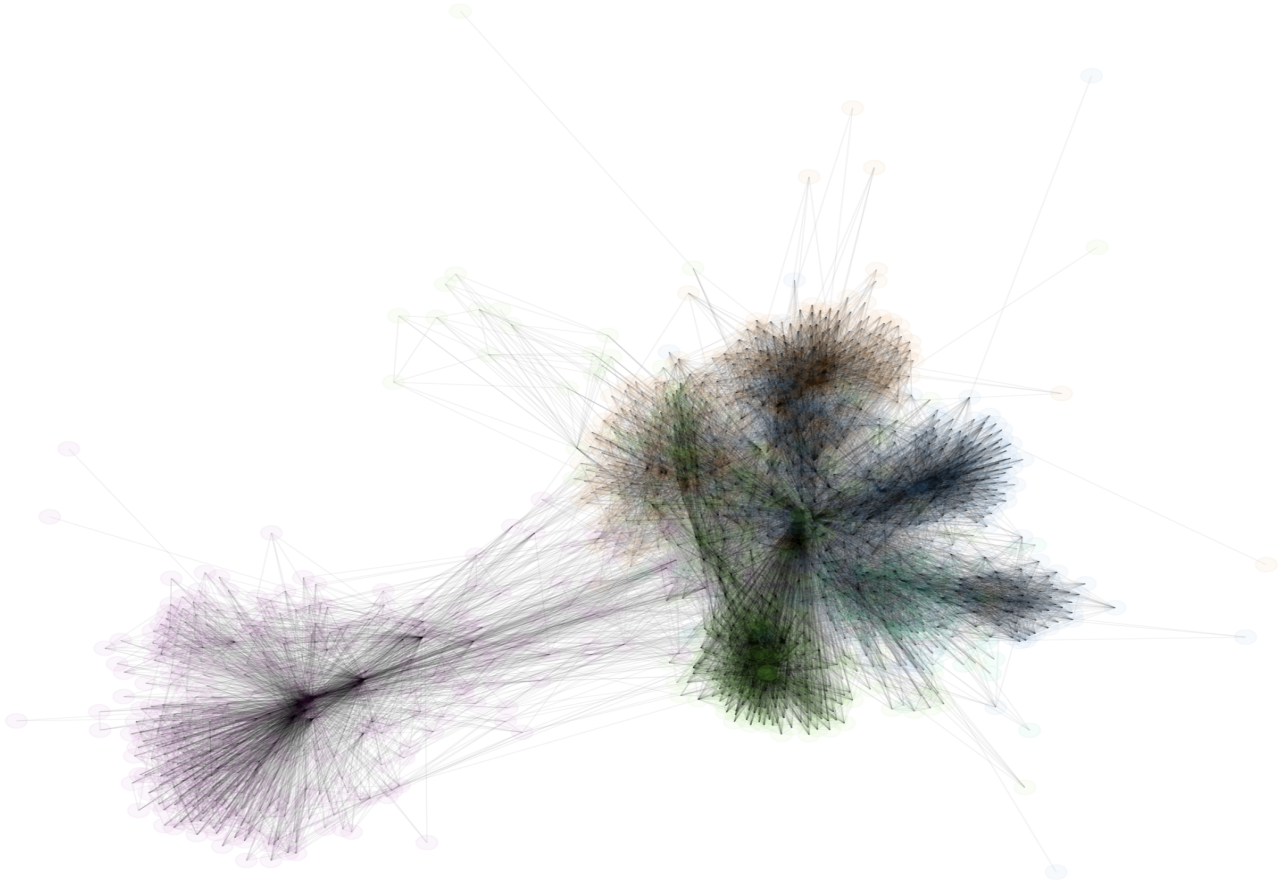
subsets based on these communities. This function uses Clauset-Newman-Moore greedy modularity maximization to find the community partition with the largest modularity. Greedy modularity maximization begins with each node in its own community and repeatedly joins the pair of communities that lead to the largest modularity until no further increase in modularity is possible (a maximum) [3][4].
By definition, modularity quantifies the community strength by comparing the fraction of edges within the community with such fraction when random connections between the nodes are made [4][6] .
Given a simple graph G with n_c disjoint communities, the modularity Q is defined as

$$Q = \sum_{c=1}^{n_c} \left[\frac{l_c}{m} - \left(\frac{d_c}{2m} \right)^2 \right]$$

where n_c is the number of clusters, l_c is the total number of edges joining nodes in a community c , and d_c is the sum of the degrees of the nodes of c . Modularity is a scalar value in the range $(-1, 1)$, with larger values implying better clustering.

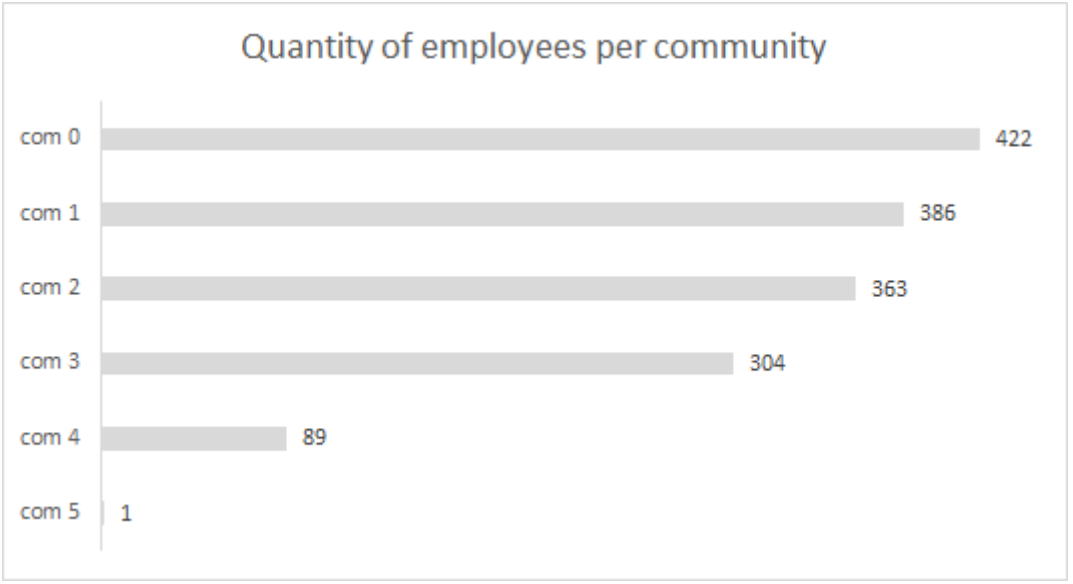
Several techniques have been utilized to investigate community structures of network, however for these project we have chosen greedy modularity to detect communities because, comparing with other methods, it has better performance to handle with large dataset and the calculations using modularity are more precise and optimal. Girvan-Newman strategy, for example, is not appropriated for our study because it has high computational complexity (i.e., $O(n^3)$, where n is the number of nodes in the community) this make it inappropriate for a large network.

To be able to apply the communities' method, we created an undirected graph, where nodes represent id_sender and id_receiver and the edges represent the messages exchanged between them. Finally, the algorithm detected 6 different communities. In the graph below, the different colors represent the 6 communities found into the network.



Social network with Communities with colors

The graph below show the quantity of employees that are on each community.



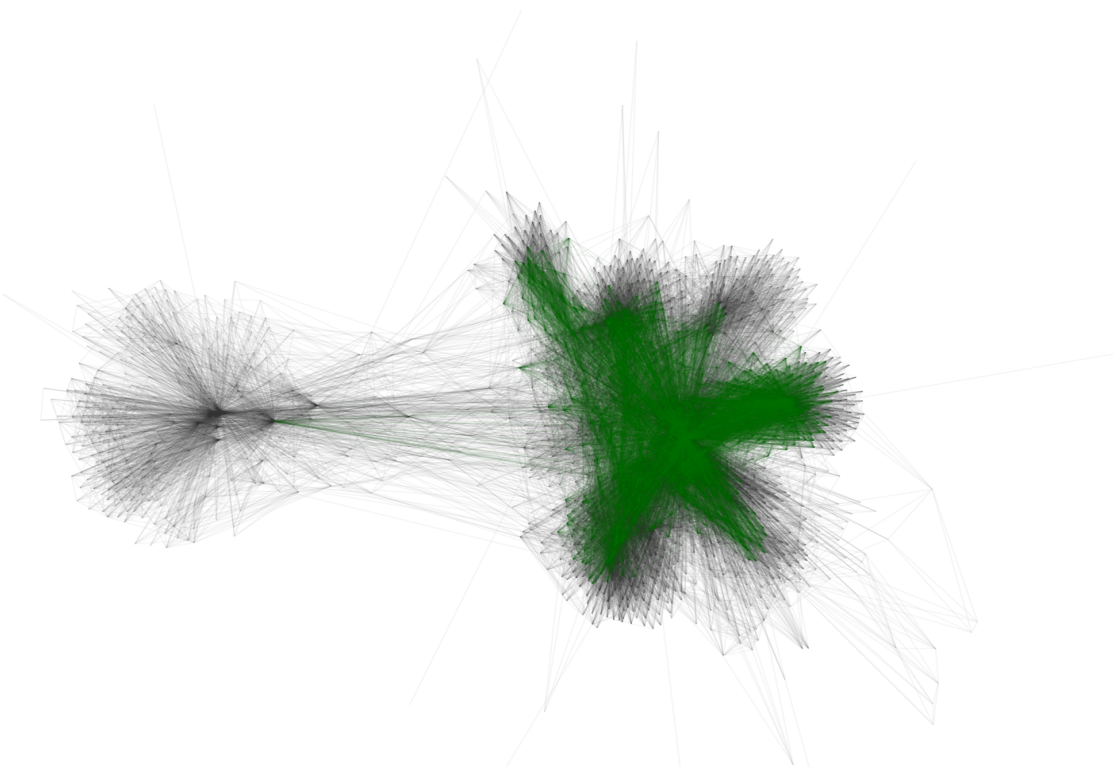
density into communities

We can suppose that community 0 is the strong community, i.e. it's the community such that each node has more neighbors inside it than in any other community [6]. Supposing that each community represent a department of the company, we can say that department 0 is one that

have more exchange messages into the organization. Possibly it's this department with an important role in the company. However, measures to optimize the use of emails between then should be tried.

b. K-Core Analysis

Given an undirected graph G, the k-core is the maximal subgraph of G in which every vertex is adjacent to at least k vertices. The most commonly used algorithm to perform -core decomposition is a pruning process that to recursively remove the nodes that have degrees less than k [5] . In our analyses, we would like to have a better visualization of groups with at least 30 degrees, what we consider already a high value of degree. Our idea was to highlight what are the most influential group of communities that we have found previously. The graph below show the result, where the green area represent the region where nodes have at least 30 degrees. In this case, it represents employees that have exchange at least 30 messages between them.



Social network with at least 30 degrees green

We suppose that the 6 different communities represent 6 possible different departments of the companies, but we don't have the department information in the data set to help us to confirm it.

c. Centrality measures

To go deeper into the communities and give resources about which are the employees are more ‘ important’, we decided to calculate the centrality measures into each community.
Each table below represent the 5 most important employees into each community detected by the algorithm, for each centrality measure.

1. Degree centrality

We suppose that the employees with the highest degree of centrality are these one that have more exchange of emails. Degree centrality is classically defined as the number of links incident upon a node, and the higher the links, the higher is the ranking of a node. From here, we can find the five most influential nodes in each cluster. As we can see below, most of these important nodes are Manager, Senior, Senior Manager or Partner, which is well justified given an employee in a higher role will have more mails exchanged due to more responsibilities.

higest degree											
community 0		community 1		community 2		community 3		community 4		community 5	
id	role	id	role	id	role	id	role	id	role	id	role
1620429	Partner	1629	Director	1331	Partner	1428	Partner	1405828	Partner	1774929	Staff
13338	Staff	1683	Assistant Manager	865	Manager	1513	Manager	1677	Partner		
1558135	Senior	940	no value	979304	Manager	1499	Senior Manager	1316	Director		
1758	Staff	13269	no value	737	Senior	1171010	Senior	1179	Senior Manager		
1238	Staff	13342	Staff	709	Senior	16092	Senior Manager	1180	Senior Manager		

2. Eigenvalue centrality

Eigenvector centrality is a measure of the influence of a node in a network. It assigns relative scores to all nodes in the network based on the concept that connections to high-scoring nodes contribute more to the score of the node in question than equal connections to low-scoring nodes. In this case, the common connections to high scoring node (people in higher roles) are also the people in authoritative positions, this can be confirmed from the table below.

higest eigenvalue											
community 0		community 1		community 2		community 3		community 4		community 5	
id	role	id	role	id	role	id	role	id	role	id	role
1620429	Partner	1629	Director	1441	Administrative Staff	1428	Partner	1405828	Partner	1774929	Staff
13338	Staff	1683	Assistant Manager	1124	Director	1171010	Senior	1179	Senior Manager		
1558135	Senior	940	no value	853	Senior	1513	Manager	1677	Partner		
1178	Administrative Staff	13269	no value	1886	Administrative Staff	1499	Senior Manager	1316	Director		
1758	Staff	1056	Partner	1273	Partner	16092	Senior Manager	1905	Partner		

3. Closeness centrality

Closeness centrality of a node is the average length of the shortest path between the node and all other nodes in the graph. Thus, the more central a node is, the closer it is to all other nodes. In this case, the center is more likely to be a role which communicates with every other department. We can observe that such a department, like administration, have more occurrences in the table below as compared to the table of degree centrality, which had more powerful roles.

higest Closeness centrality											
community 0		community 1		community 2		community 3		community 4		community 5	
id	role	id	role	id	role	id	role	id	role	id	role
1620429	Partner	1629	Director	1331	Partner	1428	Partner	1405828	Partner	1774929	Staff
1178	Administrative Staff	1683	Assistant Manager	1886	Administrative Staff	16092	Senior Manager	1609	Manager		
13468	Partner	940	no value	1124	Director	1513	Manager	1620423	Manager		
1815	Administrative Staff	885	Senior Manager	694	Staff	1501	Administrative Staff	1677	Partner		
1354	Partner	1299	Manager	880	Senior	1580	Administrative Staff	1316	Director		

4. Betweenness centrality

Betweenness centrality quantifies the number of times a node acts as a bridge along the shortest path between two other nodes. It was introduced as a measure for quantifying the control of a human on the communication between other humans in a social network. These will be the people who act as a bridge between two people of different roles. For example, a manager, between staff and partner/director.

higest betweenness centrality											
community 0		community 1		community 2		community 3		community 4		community 5	
id	role	id	role	id	role	id	role	id	role	id	role
1620429	Partner	1629	Director	1331	Partner	16092	Senior Manager	1405828	Partner	1774929	Staff
1178	Administrative Staff	1683	Assistant Manager	865	Manager	1428	Partner	1609	Manager		
13338	Staff	940	no value	979304	Manager	1513	Manager	1677	Partner		
13468	Partner	13269	no value	13215	Staff	1580	Administrative Staff	1888	Partner		
1815	Administrative Staff	1920	Senior	694	Staff	1520	Senior	1316	Director		

5. CONCLUSION

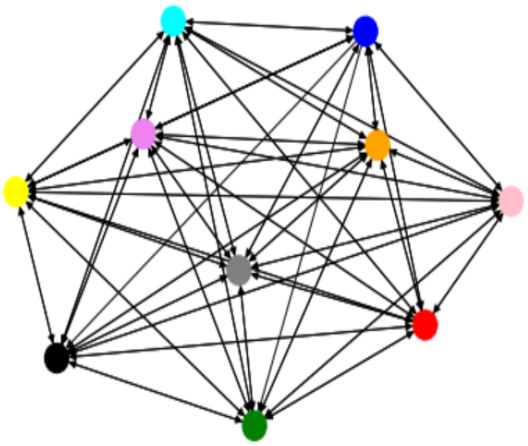
- One-on-one conversations: Out of all the 389.179 mails received, by the analysis we have identified that there are 177.150 mails which are sent to only one person, i.e., if that mail is substituted by some personal chat or an in-person conversation, the average number of mails received by a person will be (389179 -177150)/1606 = 132.02. This is a decrease by **45 %.**
- We can also advise our employees to send only extremely important emails outside of working hours (other than 0900 to 1800 hrs on working days). Such timings constitute a huge amount of irrelevant mails. We have already included an analysis on this above.
- In our previous section, we can see that clustering help us observe the possible teams in the organization, based on the assumption that a team will exchange more number of emails within themselves. Here, we observe five teams, and we've reported the five most important people in each such team.

  If we advise these teams to form a private channel or group on another platform and use that for their team related discussions, then it can really help reduce the average emails received by the employees, given all their respective teams will be discussing the work on another platform, and it will not mix up with other work related emails.

  [We cannot report the exact number by which the average mails received would decrease, but considering an average case where a cluster has some 280 employees, each receive 40% of emails from their respective teams, this would be 135.520 emails, i.e. an average decrease of **34.8 %** for all employees].

- The graph below signifies the inter-departmental communications, i.e., each node signifies one role. The color coded table below depicts the role corresponding to the color and internal emails exchanged between them.

Role/Department	Mails exchanged within the department
Administrative Staff	7.327
Assistant Manager	484
Director	6.698
Junior Staff	2.214
Staff	32.503
Manager	15.519
Senior	18.638
Partner	15.579
No Title	409
Senior Manager	9.865



Inter-Departmental Communication

The table above describes the emails exchanged within these roles. If we suggest the company to make separate groups on another platform for each of these roles, and use that instead of communicating via emails, then we can calculate the percentage decrease by subtracting the sum of the mails (=109236) in the table above from the total mails received. This decreases the average mails received by each employee by **28.06 %**.

Comments: It should be interesting to have other internal information in the data set, such as department. It could be useful to analyze, for example, if employees in some divisions mostly communicate with employees in other divisions or into theirs divisions.

6. FURTHER RECOMMENDATIONS:

1. **Virtual work teams software programs to reduce exchange of emails**. A virtual team is a group of people working across time and space and organizational boundaries using technology to communicate and collaborate. It allows having a more fast and efficient message transfer with immediate feedback [1] . Also, company reduce cost with data storage of emails and optimize communication.
2. **Promote face to face communication between people in the same project or in the same team**. If we refer our conclusions, it is clear that if communication is on one-to-one basis, maybe in person or personal chat, it is always more effective. It not only reduces the mails received by either of the parties but provide a better communication. The lengthier emails are more likely to be skipped by the person.
3. **Schedule 30 to 60 minutes to answer emails during the day**. As showed in Graph-2, employees communicate more during work hours. It should be interesting to create a policy to answer the important emails during the morning and keep afternoon to focus on activities, specially in the group of staffs. The measure can be useful to optimize the time of work.
4. **Use of AI and NLP techniques to automatically identify email based on its importance and its urgency, without employees intervention**. Reducing this manual effort to sort emails will save hours of employees, letting them concentrate on handling complex issues, and thus will boost productivity.
5. **Internal RH politics to encourage a better use of email's exchange between employees [7].**

7. REFERENCES

1. Sebastian Palus, Piotr Bródka1, Przemysław Kazienko. *Evaluation of Organization Structure based on Email Interactions.*

2. *https://en.wikipedia.org/wiki/K-means_clustering*.

3. https://networkx.org/

4. Ahmed F. Almukhtar, Eman Salih Al-Shamery (2018). *Greedy Modularity Graph Clustering for Community Detection of Large Co-Authorship Network.*

5. i-Xiu Kong, Gui-Yuan Shi, Rui-Jie Wu, Yi-Cheng Zhang (2019). *k-core: Theories and applications*

6. Filippo Menczer, Santo Fortunato, Clayton A. Davis (2020). *A First Course in Network Science.*

7. Mano, Rita & Mesch, Gustavo. (2012). *E-mail and work performance*. Encyclopedia of Cyber Behavior. 1. 106-116. 10.4018/978-1-4666-0315-8.ch009.