Written Report – 6.419x Module 3

Name: (luis_go95)

2. Problem 1

Part (c): (2 points) How does the time complexity of your solution involving matrix multiplication in part (a) compare to your friend's algorithm?

Solution:

In time complexity it can be always as high or more complex than the previous one but if the matrix is so big, it will probably run a big quantity of empty relations. In contrast, the previous versions only run in real relations.

Part (d): (3 points) Bibliographic coupling and cocitation can both be taken as an indicator that papers deal with related material. However, they can in practice give noticeably different results. Why? Which measure is more appropriate as an indicator for similarity between papers?

Solution:

It is truth that study similar material but the way in how to do it is different. Bibliographic coupling literally means the frequency of when two papers cite the same citation, meanwhile the co-citation refers to the frequency of when two papers are cited together. As a result, the first one is going to be a better metric in the case when you want to study papers that follow the same methodology or stuff like that. In contrast, the second one is more appropriated when you are more interested in the content of the paper or the evolution of a specific topic. In conclusion, the co-citation network is going to be more appropriate because it is going to capture a better relationship if we are interested in the true content of a paper.

3. Problem 2

Part (c): (2 points) Observe the plot you made in Part (a) Question 1. The number of nodes increases sharply over the first few phases then levels out. Comment on what you think may be causing this effect. Based on your answer, should you adjust your conclusions in Part (b) Question 5?

Solution:

Probably in the first 3 months the logistics and sales improved and therefore there was the need for more people to participate. After the first seizure in the fourth month, and because of the reorientation of bringing cocaine from Morocco, and then from Colombia, probably many connections needed to be changed or removed. As a result, I would change the conclusion of the part b question 5 but by separating the average results, one result only considering the months when they imported from Morrocco, and other results considering the months when they imported from Colombia

Part (d): (5 points) In the context of criminal networks, what would each of these metrics (including degree betweenness, and eigenvector centrality) teach you about the importance of an actor's role in the traffic? In your own words, could you explain the limitations of degree centrality? In your *opinion, which one would be most relevant to identify who is running the illegal activities of the group? Please justify.*

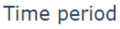
Solution:

All the nodes that have a high score in betweenness centrality are key nodes that help as important connections between different parts of the organization, and therefore are key for information transition. They participate in a more quantity of shortest paths between all nodes, and therefore are important nodes that if you do not have them, probably the flow of the organization is going to be heavily affected. All nodes that have a high score in eigenvector centrality are the most important nodes in the since that have an important participation in the graph but also, are related with other nodes that also are so important. Probably these nodescan be the leaders of the organization. If you cut them, the graphs is going to be affected strongly. All nodes that have high score in degree centrality means that are the most popular ones because they interact with a bigger quantity of other nodes. However, degree centrality have some limitations like: (1) does not consider the weights of the edges, so a node with high degree can be connected to many people but not necessary be important, (2) only consider the immediate neighbors, for example, it can be that a node with high degree have many connections but the connections of its connections are not so important or are few, therefore measures like eigenvector centrality can be better in this sense.

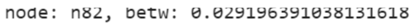
In my opinion, the ones that are running the illegal activities are those with high values in betweenness centrality because they are the key nodes for information transitionand probably, the leaders of logistics. If you cut them, the graph is going to received an important impact. In contrast, those with high eigenvector centrality are more like the leaders of the organization and not necessarily those responsible in that the things happened. However, of course, if you cut the heads, the impact also is going to be big

Part (e): In real life, the police need to effectively use all the information they have gathered, to identify who is responsible for running the illegal activities of the group. Armed with a qualitative understanding of the centrality metrics from Part (d) and the quantitative analysis from part Part (b) Question 5, integrate and interpret the information you have to identify which players were most central (or important) to the operation.

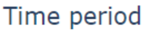
In my opinion, a person is truly important to the organization if have one of these characteristics: (1) This person is in a key position in the sense that connects groups of nodes to other groups, and as a result, are key nodes for information sharingand also probably are responsible of that everything happened in the organization. If we cut one of them, the impact will be big in the sense of affecting the logistics of the group. (2) This person is part of the heads of the organization. This is because if we cut some of them, the impact will be big in order that there is not temporally someone that takes important decisions. The first characteristic is related with betweenness centrality and the second one with eigenvector centrality.



Top 10 average betweenness centrality



```
node: n82, betw: 0.029196391038131618
```



Top 10 eigenvector centrality

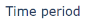
```
node: n9, eig: 0.10068037663051316
```

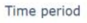
In the previous graphs and images we can see the top 10 of most important nodes considering eigenvector and betweenness centrality as means of all the periods and their values over time. From this, in my opinion, the most important nodes of the organizations are nodes 1 and 3 because they are in the top 3 of the highest scores in both centrality measures. This means that if we cut at least 1 of them, the impact of the organization is going to be catastrophic. However, there are some other important nodes that also are very important like n12, n76 and n87 if we analyze betweenness centrality, and n85, n76 and n83 if we analyze eigenvector centrality.

Part (f) Question 2: (3 points) The change in the network from Phase X to X+1 coincides with a major event that took place during the actual investigation. Identify the event and explain how the change in centrality rankings and visual patterns, observed in the network plots above, relates to said event.

Solution:

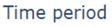
According to the problem description, the first seizure happened in Phase 4. As a result, traffickers reoriented to cocaine import from Colombia. We analyze how this situation impacted the nodes centrality measures scores:



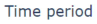


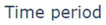
Visually, the most important changes that we can mention between phase 4 and 5 are (1) that the node "n12" suddenly have an increase both in betweenness and eigenvector centrality, and (2) nodes "n89", "n", "n83", "n9" suddenly have a decrease in both centrality measures. In the next graph we can observe it easily:

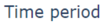
Increase



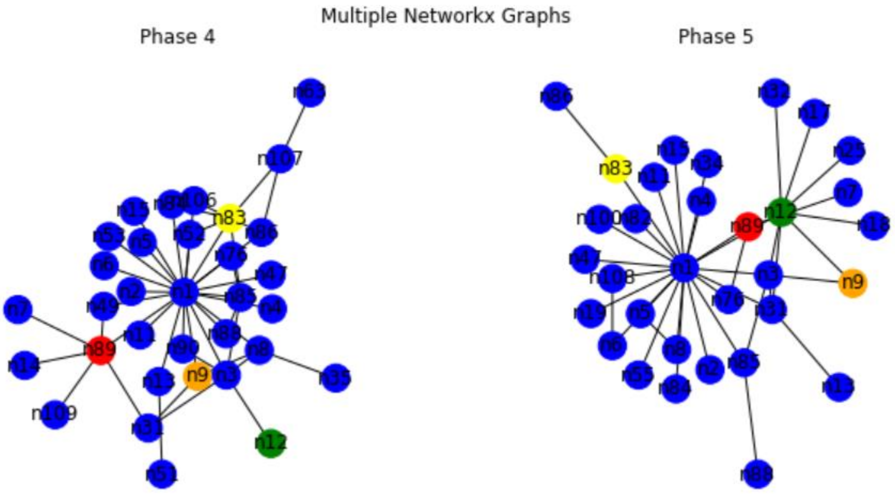
Decrease







This relates with the event because probably "n89", "n", "n83", "n9" were nodes that could have had an important participation when the logistic was importing from Morrocco. In contrast, probably node "n12" in those periods could have taken a bigger participation when importing from Colombia. Finally, in next graph, we can see how these nodes change their participation in the graph:



Part (g): While centrality helps explain the evolution of every player's role individually, we need to explore the global trends and incidents in the story in order to understand the behavior of the criminal enterprise.

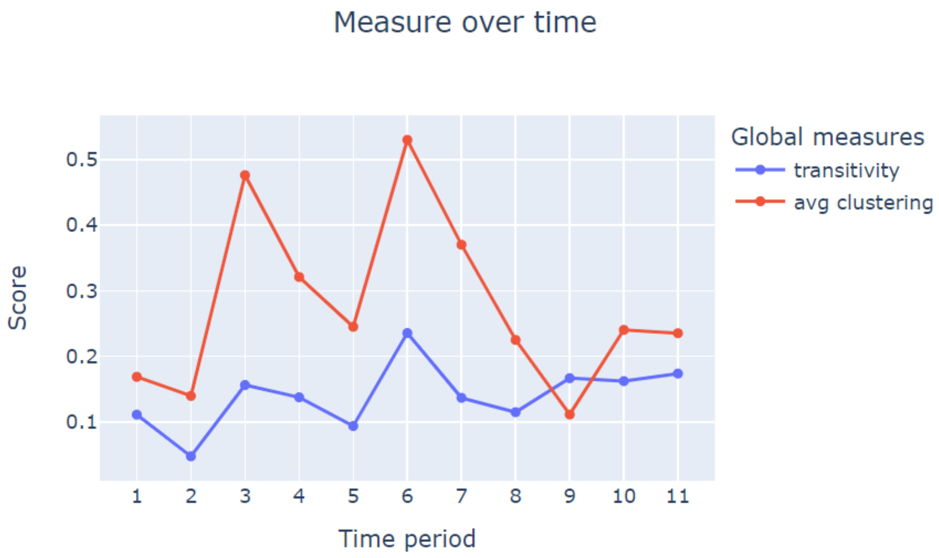
Describe the coarse pattern(s) you observe as the network evolves through the phases. Does the network evolution reflect the background story?

Solution:

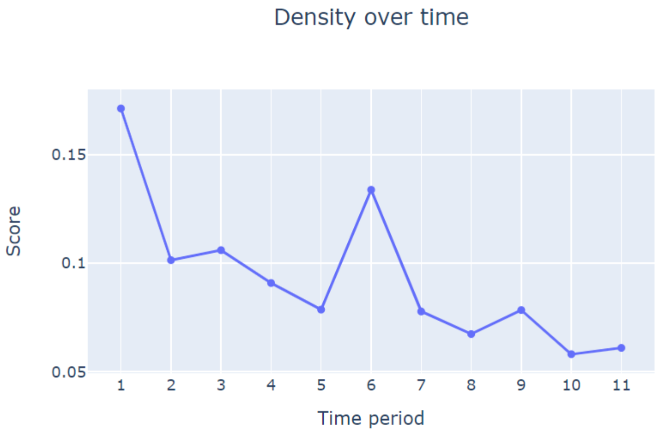
Considering all the seizures over time, we can analyze and compare how it relates with our statistics.

Phase 4	1 seizure	$2,500,000	300 kg of marijuana
Phase 6	3 seizures	$1,300,000	2 x 15 kg of marijuana + 1 x 2 kg of cocaine
Phase 7	1 seizure	$3,500,000	401 kg of marijuana
Phase 8	1 seizure	$360,000	9 kg of cocaine
Phase 9	2 seizures	$4,300,000	2 kg of cocaine + 1 x 500 kg marijuana
Phase 10	1 seizure	$18,700,000	2200 kg of marijuana
Phase 11	2 seizures	$1,300,000	12 kg of cocaine + 11 kg of cocaine

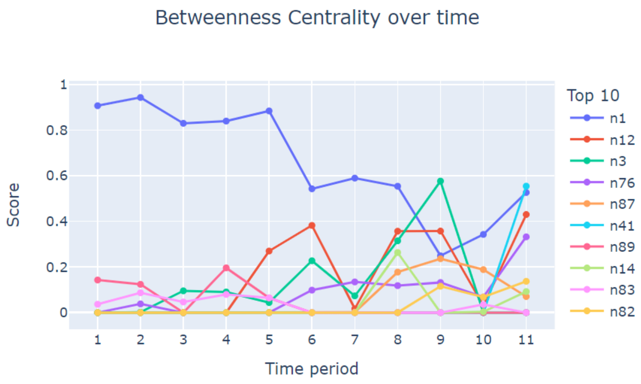
First of all, we can analyze them looking for global measures like transitivity, avg clustering or density.



In the previous graph we can see for example phase 4 that is related when the first seizure was made and also the import from Morrocco changed to import from Colombia. As a result, we can visualize a decrease in transitivity and avg clustering which means that the network was importantly affected, and they lose a little the well working organization. Another fact is around phase 7 where also it was a really important seizure of marijuana (401 kg) and therefore the metrics during this period started to decrease again.



Also, if we compare the seizures with the density score over time, we can see that overall, the network has a tendency of decreasing. This is totally in concordance with the seizures and also, we can see an important decrease from phase 7 and so on because in next phases, the quantity of drugs seizures increased considerably.



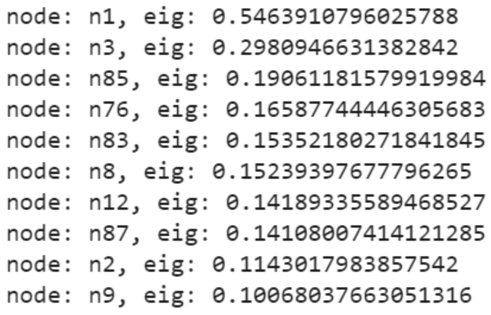
If we analyze again centrality measures like betweenness over time, we can contrast these facts. For example, since phase 4, we can see that “n12” started to increase their important and is totally in concordance because this node is Ernesto Morales, the principal organizer to import from Colombia. At the same time, we can see that in phase 7 their score decreases a lot because of the big seizure of marijuana. Also from the facts, we know that in phase 10, there was the most important seizure (2200 kg of marijuana) and in the same graph, we can see how all the important nodes were really affected, between them we can mention “n3” Piere Perlini (Principal of Serero), and “n12” Ernesto Morales again, but “n1” Daniel Serero had a little increase.

Part (h): Are there other actors that play an important role but are not on the list of investigation (i.e., actors who are not among the 23 listed above)? List them, and explain why they are important.

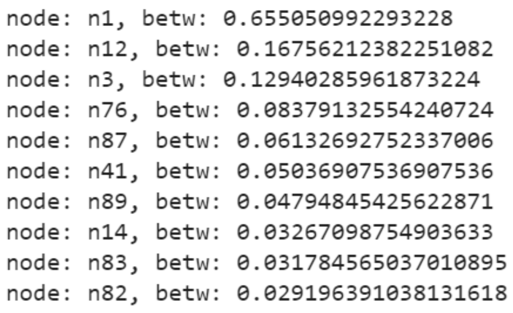
Solution:

Considering all the seizures over time, we can analyze and compare how it relates with our statistics.

Top 10 of eigenvector and betweenness centrality



```
node: n1, eig: 0.5463910796025788
node: n3, eig: 0.2980946631382842
node: n85, eig: 0.19061181579919984
node: n76, eig: 0.16587744446305683
node: n83, eig: 0.15352180271841845
node: n8, eig: 0.15239397677796265
node: n12, eig: 0.14189335589468527
node: n87, eig: 0.14108007414121285
node: n2, eig: 0.1143017983857542
node: n9, eig: 0.10068037663051316
```



```
node: n1, betw: 0.655050992293228
node: n12, betw: 0.16756212382251082
node: n3, betw: 0.12940285961873224
node: n76, betw: 0.08379132554240724
node: n87, betw: 0.06132692752337006
node: n41, betw: 0.05036907536907536
node: n89, betw: 0.04794845425622871
node: n14, betw: 0.03267098754903633
node: n83, betw: 0.031784565037010895
node: n82, betw: 0.029196391038131618
```

If we compare to the top 10 of higher scores in previous centrality measures, we can see that there are nodes that are not considered in the 23 list. In the case of eigenvector centrality, we have nodes “n2” and “n9”, and in the case of betweenness centrality we have nodes “n41” and “n14”. If these nodes have a high score in their metrics, it means that they are really important to the organization, and we probably need to take them into account also.

Part (i): What are the advantages of looking at the directed version vs. undirected version of the criminal network?

Solution:

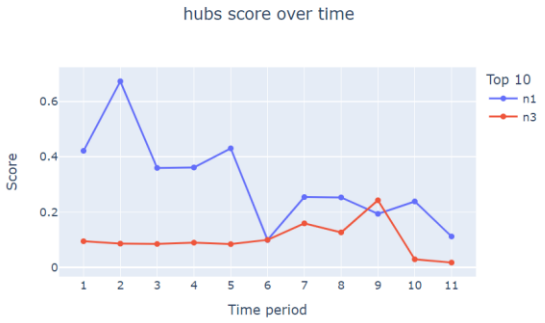
Performing a directed version can give us some advantages. In the case of degree centrality, with the in degree, we learn how many nodes are directly connected to the current node, in the case of out degree, how many nodes my current node connects to. Probably in our context, it could be interesting for us to know who are those nodes that have high score both in in and out degree because this means that are nodes that connects to many nodes to many mores, therefore can be important in order to impact the drug organization. Also, we can make the comparison between left – right eigenvector centrality. In the first one, we learn that those with high left eigenvector centrality means that the current node transmits information to probably many important nodes. In the case of right eigenvector centrality, a high score means that the node receives information from many important nodes. In our context, it can be very helpful because probably if we arrest those with high left eigenvector centrality, the graph is going to be highly impacted because those more important for information transmition are not going to be anymore.

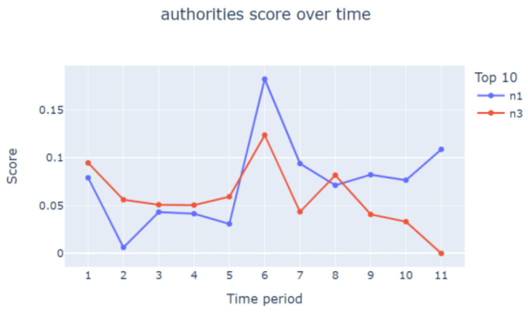
Part (j): (4 points) Recall the definition of hubs and authorities. Compute the hub and authority score of each actor, and for each phase. (Remember to load the adjacency data again this time using create_using = nx.DiGraph().)

With networkx you can use the nx.algorithms.link_analysis.hits function, set max_iter=1000000 for best results.

Using this, what relevant observations can you make on how the relationship between n1 and n3 evolves over the phases. Can you make comparisons to your results in Part (g)?

Solution:





As a first impression, we can see that in the Hubs score over time, “n1” tends to have the higher score and it is totally coherent because Daniel Serero is the master mind of the organization and therefore, he is on charge of take the important decisions and order to others. Also, in the case of the authorities score over time, “n3” Pierre Pierlini is in many phases the higher score and is coherent because is the principal of “n1”, therefore is the first one in receive the most important information.

The most remarkable comparison that we can make from the seizures through phases is the decrease that we can see in the authorities graph in the phase 7, 9 and 10 because in those moment happened the most important seizures. We can see that both “n1” and “n3” were impacted by an important reduction in their scores but most remarkably “n3”. “n1” in last phases shows that always got a high enough score, which is coherent because of being the leader of the organization. In other words, in times of crisis, probable “n1” needed to participate more actively and that’s because the high score in authority.

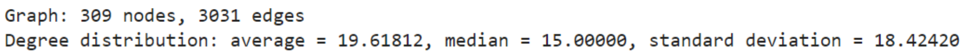
4. Project: email-EU-core Temporal Network

¿Does the interaction in Department 1 or Department 2 in a company is more organized and structured?

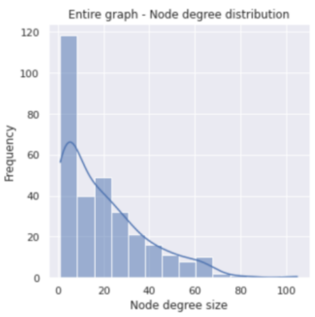
The following analysis is based on “email-EU-core Temporal Network” from public Standford datasets. It is a directed dataset where emails are only considered inside the organization and for 803 days. Also, there is compare the dataset for “email-Eu-core-temporal-Dept1.txt” vs “email-Eu-core-temporal-Dept4.txt”. The approach for analyzing the datasets will be based on showing (1) General graph information, (2) Global measures comparison, and (3) individual centrality measures comparison.

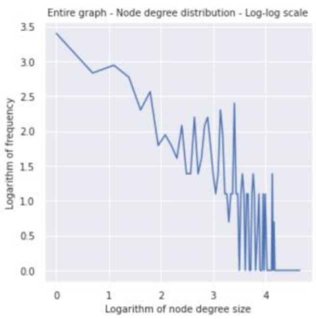
General information comparison

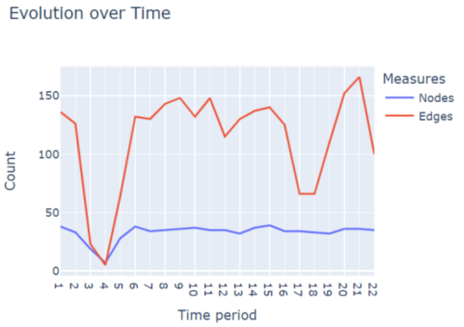
Department 1



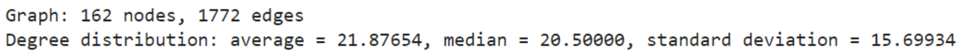
```
Graph: 309 nodes, 3031 edges
Degree distribution: average = 19.61812, median = 15.00000, standard deviation = 18.42420
```



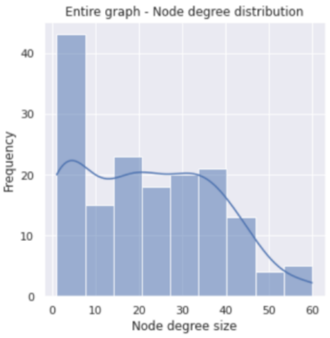


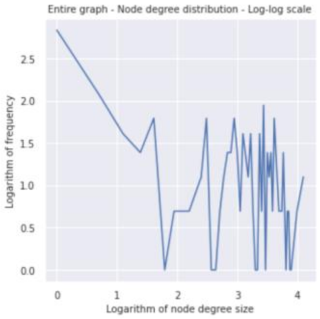


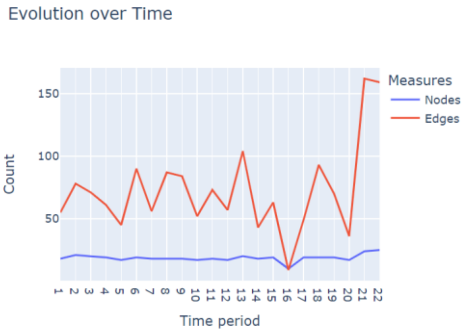
Department 2



```
Graph: 162 nodes, 1772 edges
Degree distribution: average = 21.87654, median = 20.50000, standard deviation = 15.69934
```





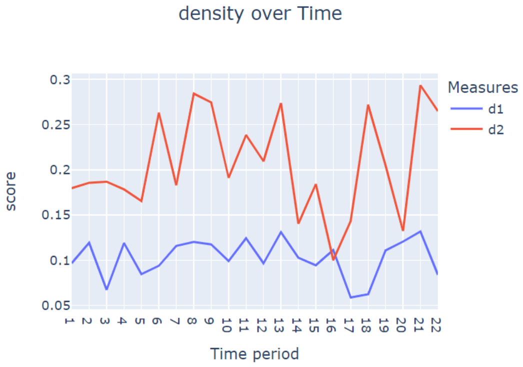


We can see from the previous graphs are that (1) both networks seem to follow a Power Law distribution, (2) Department 2 is approximate 2 times department 1 if we compare the total nodes of all the period, and (3) the quantity of edges have many ups and downs over time that can be a result of vacation periods or some other special situations in the company.

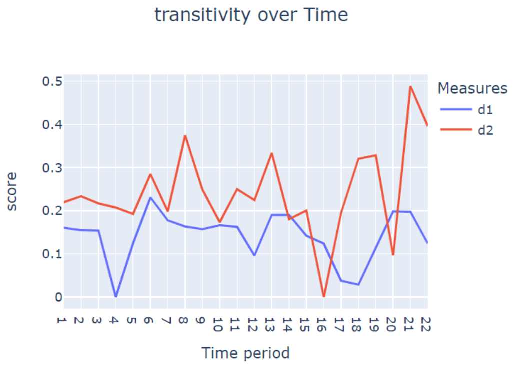
Global measures comparison

Comparison

The idea of this part was to compare each global measure (density, transitivity, average clustering and greedy communities) over time and try to determine with a Two-sample t-test if the mean of one department is higher than the other. In order to achieve that, first, it was necessary to validate if the data is normally distributed and has equal variance. For that purpose, it was used an alfa = 0.05

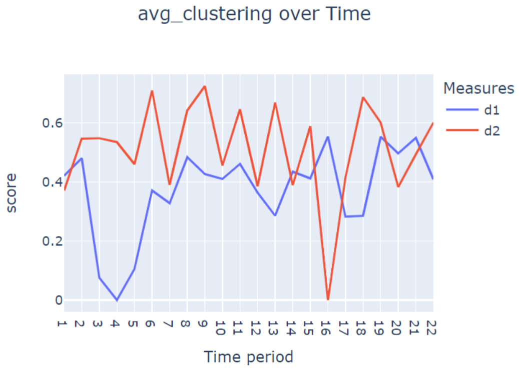


```
Shapiro-Wilk test for normality Department 1
statistic=0.926, p-value=0.101

Shapiro-Wilk test for normality Department 2
statistic=0.935, p-value=0.160

Levene's test for equality of variances
statistic=11.503, p-value=0.002
```

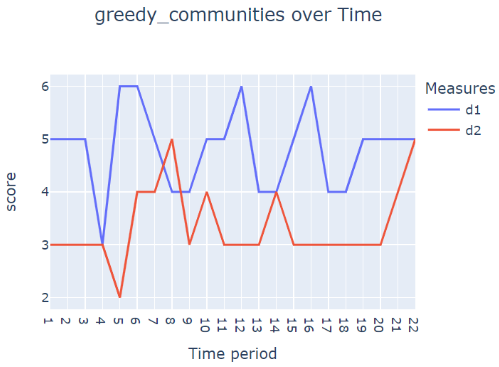


```
Shapiro-Wilk test for normality Department 1
statistic=0.904, p-value=0.035

Shapiro-Wilk test for normality Department 2
statistic=0.952, p-value=0.343

Levene's test for equality of variances
statistic=3.032, p-value=0.089
```

In the case of density, that is how well connected is a graph taking into account all the edges over all possible edges, we can see that department 2 tends to be higher. However, the data does not meet the requirements of normality, and as a result, the test was invalid to perform. In the case of transitivity, that is the proportion of all possible triangles in the network that actually exist, visually it is harder to determine if some department is better over time. However, it also does not meet the requirements both for equality of variances and normality.



```
Shapiro-Wilk test for normality Department 1
statistic=0.887, p-value=0.017

Shapiro-Wilk test for normality Department 2
statistic=0.886, p-value=0.016

Levene's test for equality of variances
statistic=0.199, p-value=0.658
```



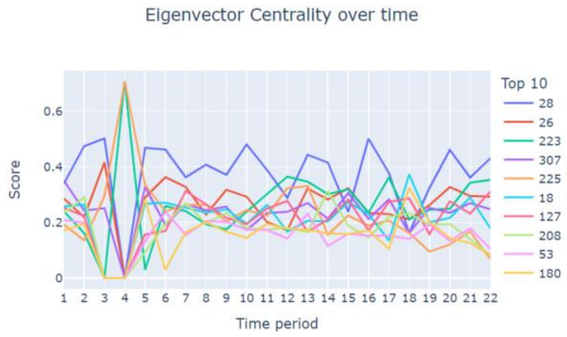
```
Shapiro-Wilk test for normality Department 1
statistic=0.862, p-value=0.006

Shapiro-Wilk test for normality Department 2
statistic=0.766, p-value=0.000

Levene's test for equality of variances
statistic=0.226, p-value=0.637
```

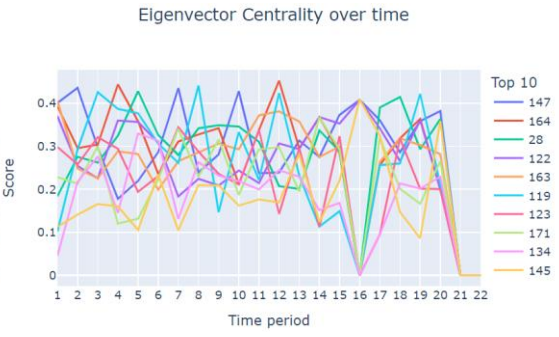
In the case of average clustering measure, that describes the degree of clustering within a network as a whole, we can see that department 2 tends to be higher, but if we run the tests, we cannot perform the two samples test because the requirement of equality of variances is not met. In the case of the greedy communities, that is a methodology for getting clusters from our network, from the visualization, we can think that department 1 tends to have more clusters. However, if also cannot run the test because equality of variances is not met.

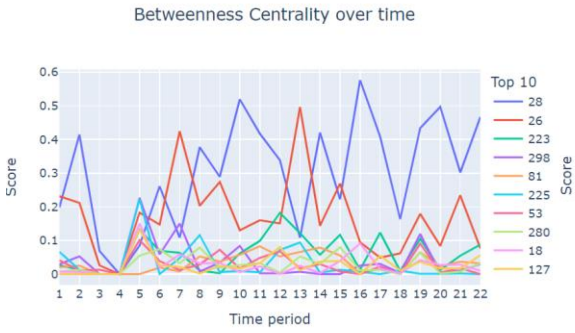
Individual measures comparison

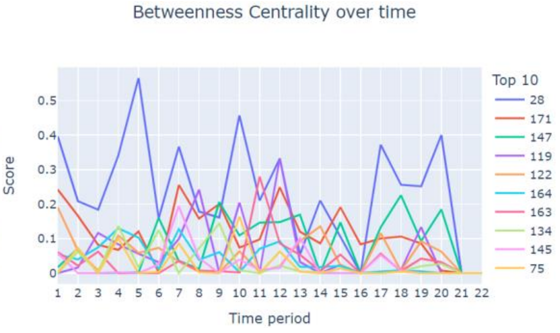
The aim of this part is to have a big picture of the centrality measures over time for the top 10 best evaluated nodes for each department

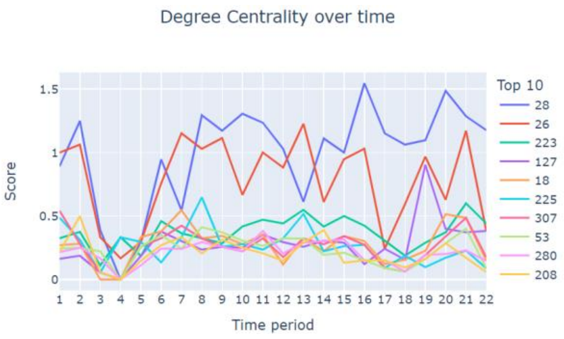
Departement 1 **Departement 2**

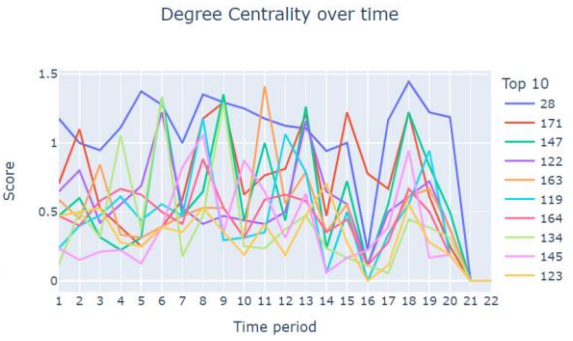


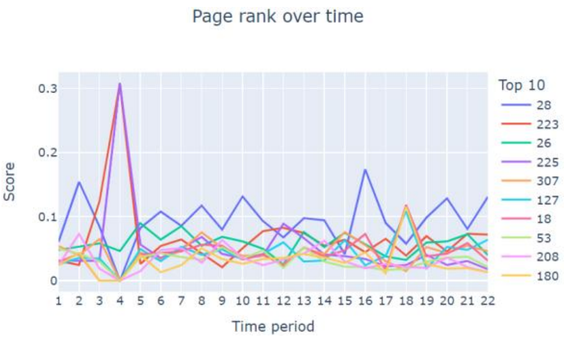


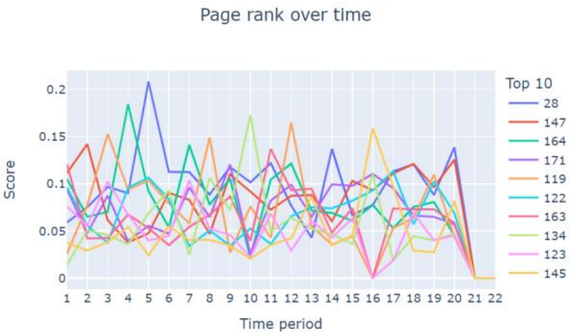


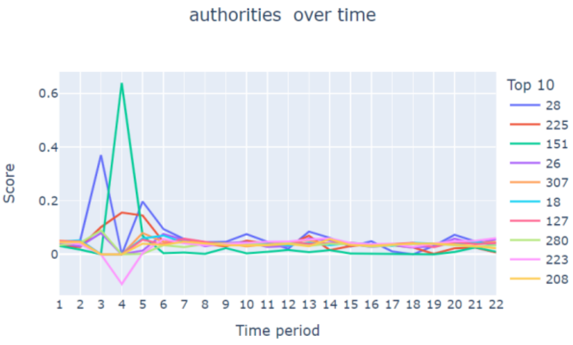


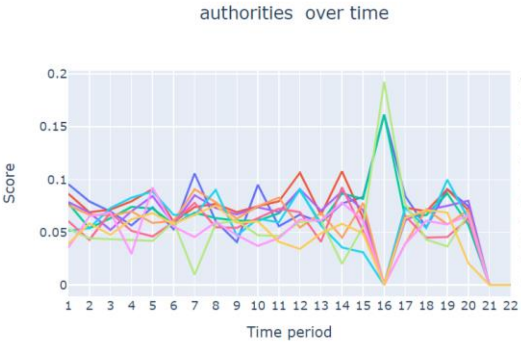


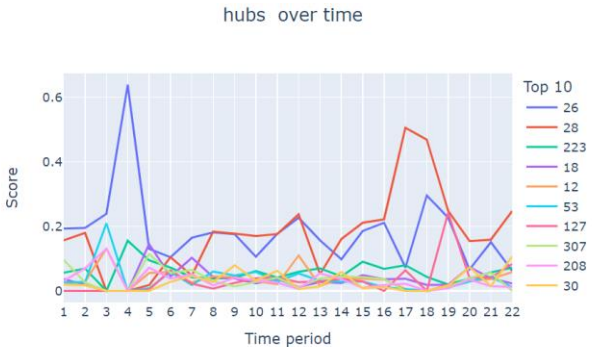


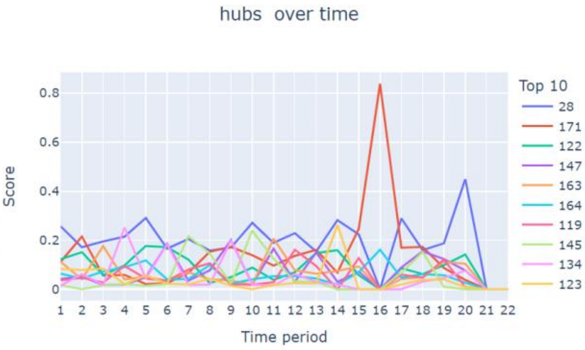


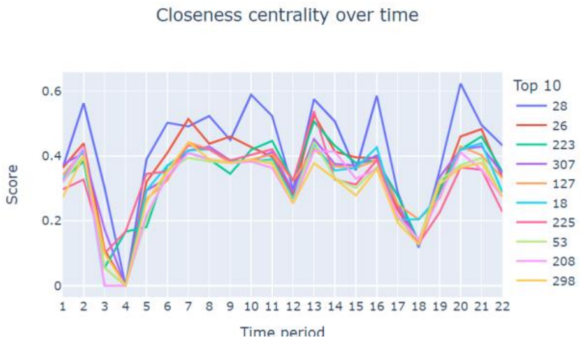


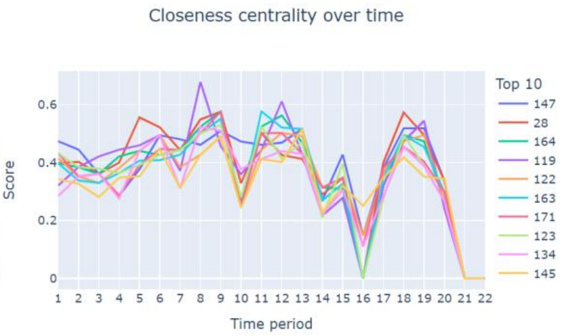


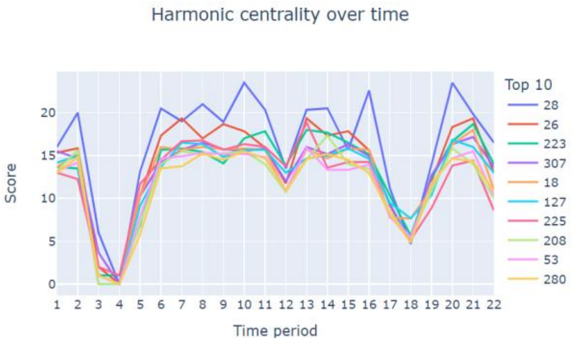


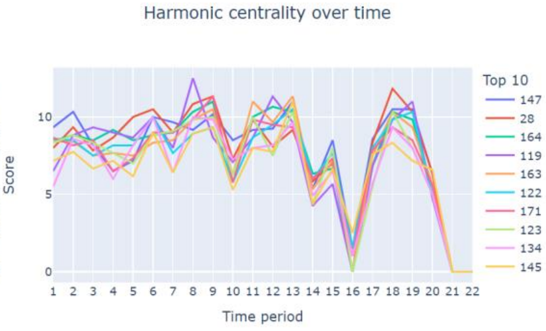








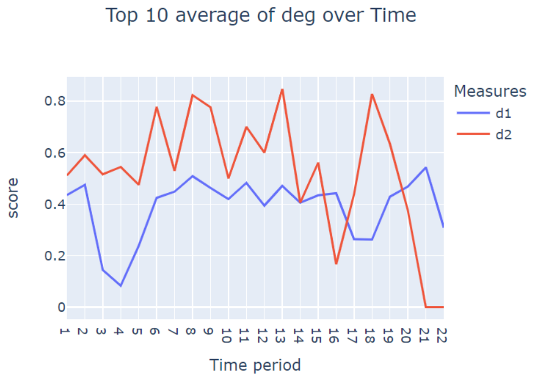




The most important observations that we can mention about previous graphs are that (1) node "28" probably is a person that has an important impact in both departments. This means that probably it has a high role in the company and therefore for some reason he appears on both networks. This person has a high score in eigenvector, betweenness, degree and page rank centrality because are measures which purpose is to determine the importance of a node. This is important because if this node is removed, the department communications would be highly impacted. Also, (2) the most important nodes in department 1 are "28", "26" and "223" because are almost always in the top 3 in most of the centrality measures including eigenvector, betweenness, degree, page rank centrality, etc. Finally, (3) some of the most important nodes in department 2 are "28", "171", "147" and "164" because in most cases are in the top 5 of the centrality measures. In both cases for department 1 and 2, if we remove some of the important nodes, the communication will have an important impact.

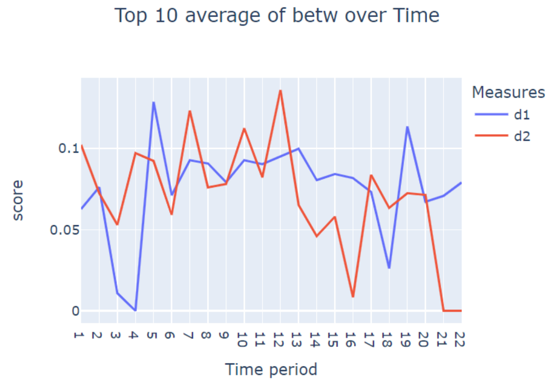
Top 10 average comparison

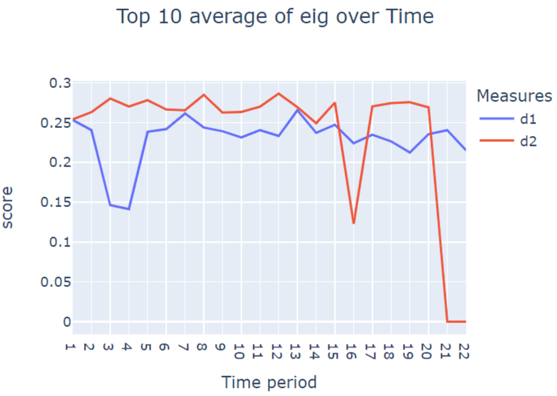
The purpose of this part was to compare the average of the top 10 best scored nodes from each department in order to determine if in average, that specific top 10 is higher in a specific department. For achieving that, it was necessary also to run normality test, equality of variances test and if that was ok, a Two sample t-test for mean comparison.

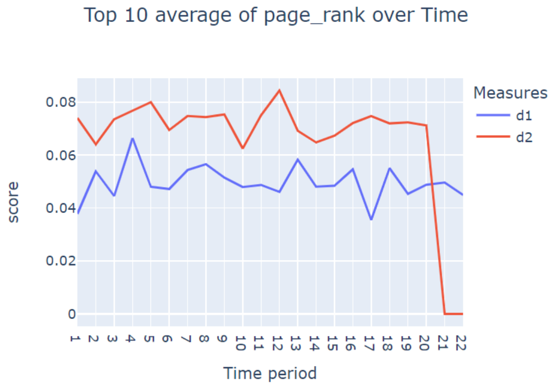


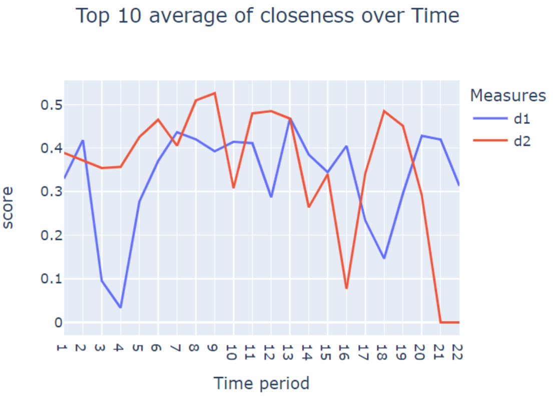
```
Shapiro-Wilk test for normality Department 1
statistic=0.854, p-value=0.004

Shapiro-Wilk test for normality Department 2
statistic=0.910, p-value=0.047

Levene's test for equality of variances
statistic=4.930, p-value=0.032

Two-sample t-test for deg mean comparison:

t-statistic:  -2.442513633783928
p-value:  0.01887379741843584
Department 2 has a higher mean than Department 1.
```

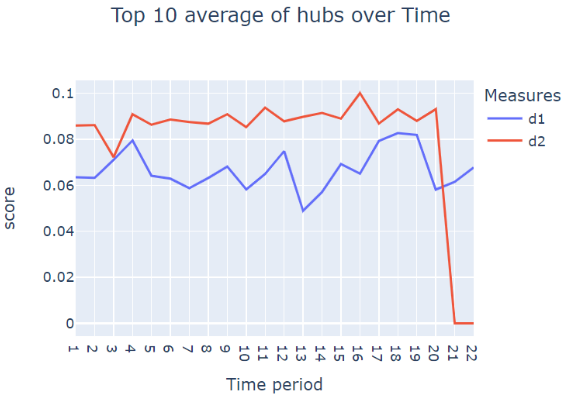
In the case of degree centrality comparison, all the requirements were meet for performing the t-test and as a result, we determined that the mean of the top 10 best scored nodes of department 2 have a higher mean that the top 10 best scored nodes of department 1.

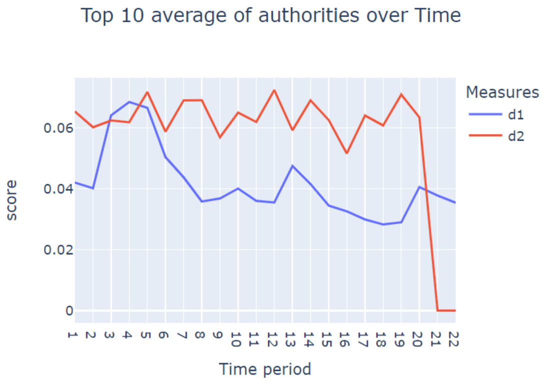


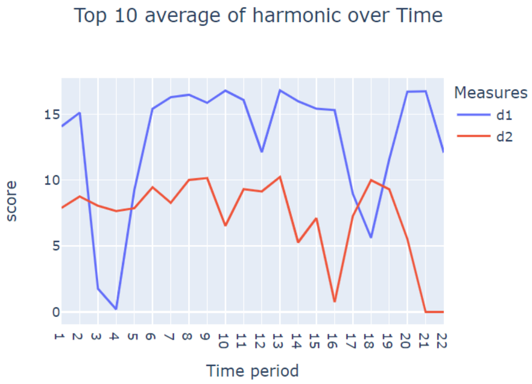







Top 10 average of hubs over Time


Top 10 average of authorities over Time


Top 10 average of harmonic over Time

For all the previous graphs, the normality and equality of variance test were performed but in any case they were met. As a result, the t-test wasn't performed. However, if we only take into account the visualization, it seems that in most cases department 2 tends to have a higher mean.

Conclusions

- Both networks seem to follow a Power Law distribution
- Department 2 is approximate 2 times department 1 if we compare the total nodes of all the period
- In almost all the global measures like density, transitivity and average clustering, department 2 tends to have a higher score. However, the data does not meet the requirement of normality and equality of variances in order to statistically validate that.
- Node "28" probably is the most important person in both departments.
- Nodes "28", "26" and "223" are the most important people in department 1
- Nodes "28", "171", "147" and "164" are the most important people in department 2
- The mean of the top 10 best scored nodes in degree centrality of department 2 have a higher mean that the top 10 best scored nodes of department 1.