Data Analytics **2023**

Names: Erifeoluwa Jamgbadi, Fabrizio Miguel Dandreamatteo

Link Analysis on Autonomous Systems

Due date: 12.04.2023

Introduction

Autonomous system is a system which contains a group of IP(Internet Protocol) routing prefixes that follows predefined policies on how information is sent and received within and outside the network. This relates to our project because our dataset gives information about a computer network system. We have a system of computers that are connected to many servers and in this system information is sent and received through these connections.

Link Analysis is used to help understand the relationship between our system which is made up of computers and servers. In this report, we formulate questions and answer these questions as we go through steps to preprocess our data and preform link analyses. We learn more about the relationship between the server and computer through the use of centralities in link analyses and displaying these graphs.

Data exploration and pre-processing

Question 1 - What does the data look like?

The dataset can be found in ’11_Autonomous_Systems.csv’. After converting the csv file into a pandas’ DataFrame, we can see that the column labled **ComputerNumber** almost match the index number. The only difference is that the indexing starts with 0, and the **ComputerNumber** column starts counting at 1.

The other 21 columns have labels named after Greek Gods. We can infer that these labels represent the servers’ names. It makes more sense to represent the DataFrame with the **ComputerNumber** as index and the servers as columns.

From what can be seen in the data frame, the value for each columns seem to be zeros (0) and ones (1). Below is an tabulated example of what the data looks like, 153 computers (rows) and 21 servers (columns).

	Zeus	Hera	Poseidon	Demeter	Athena	Apollo	...	Vulture
ComputerNumber								
1	0	0	0	0	1	0	...	0
2	0	0	0	0	1	0	...	0
3	0	0	0	1	0	0	...	0
4	0	0	0	1	0	0	...	1
5	0	0	0	0	1	0	...	1
...	...	...	...	...	...	...	...	...

Table 1: Representation of the data

Question 2 - Are there any missing values in the data set? If not, are all values binary?

Upon confirmation that all values in the columns exist and are binary, we can infer that the value one (1) represents a connection between a computer and a server and conversely, the value zero (0) shows no connection with the server. From here, we can also infer that a computer can only be connected with a server and a server can only be connected with a computer. From here on we can ask numerous questions about the data.

Question 3 - Which are the least and most connected server?

In the figure below, the most and least connected server can be observed. The server "Granite" is the most connected with 69 computers, while "Hera", "Hermes", "Shale" and "Atlas" are the least connected with only 8 computers.

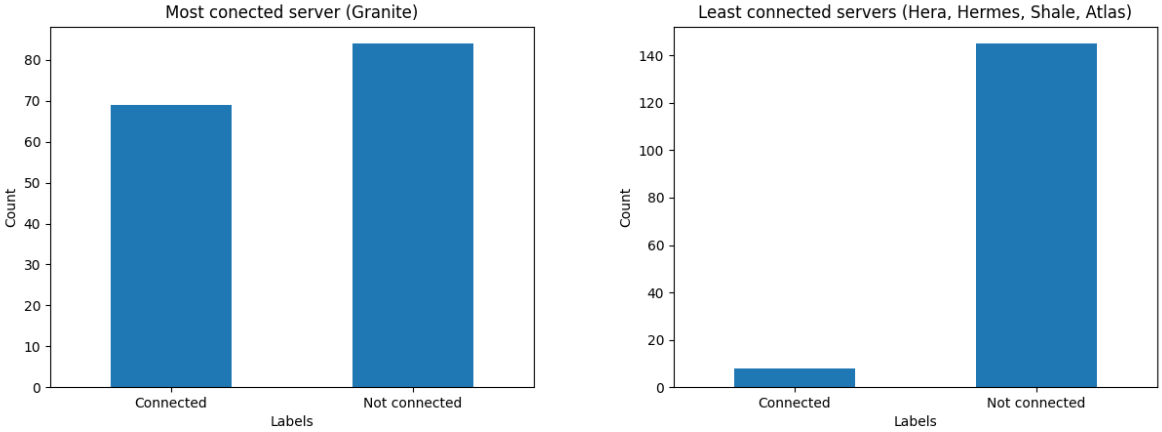


Figure 1: Most (69 computers) and least (8 computers) connected servers and servers with above average connections

Question 4 - What is the mean number of connections and which servers have more than the average?

The mean number of connections among the servers is 19 connections (rounded to the closest integer). Only 5 of 21 servers surpass the average connections. They are "Zeus" (23), "Athena" (52), "Ares" (46), "Hephaestus" (20) and "Granite" (69).

Question 5 - Are there any computers that are not connected to any server? if so, which ones?

There are 14 computers that are not connected to any server. They are: [34, 48, 49, 51, 72, 81, 99, 111, 115, 134, 135, 144, 148, 150]. In contrast, there is a maximum number of servers that computers can connect with. Computers 108 and 126 represent this maximum and they are connected to 7 different servers.

Question 6 - What does the distribution look like?

A summary visualization of all the aforementioned questions can be expressed in Figure 2. It shows in the form of a bar chart the connections of each server highlighting the maximum number of computers (blue line) and the average connection with the servers (red line). The answer to question 4 can be visually represented here with the orange bars that surpasses the red line.

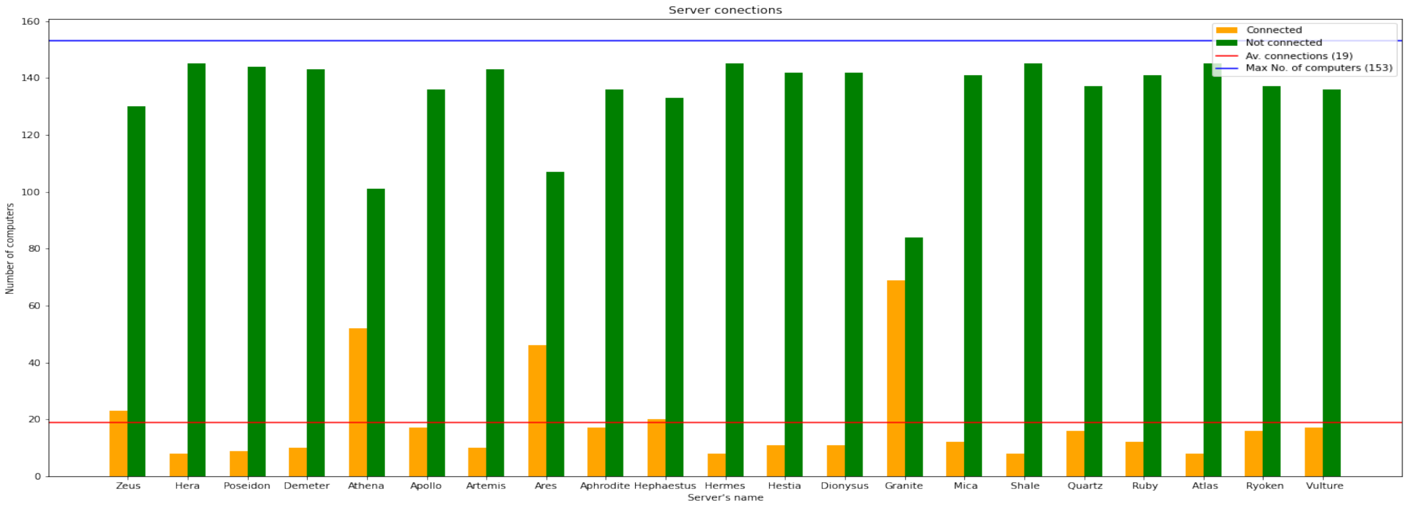
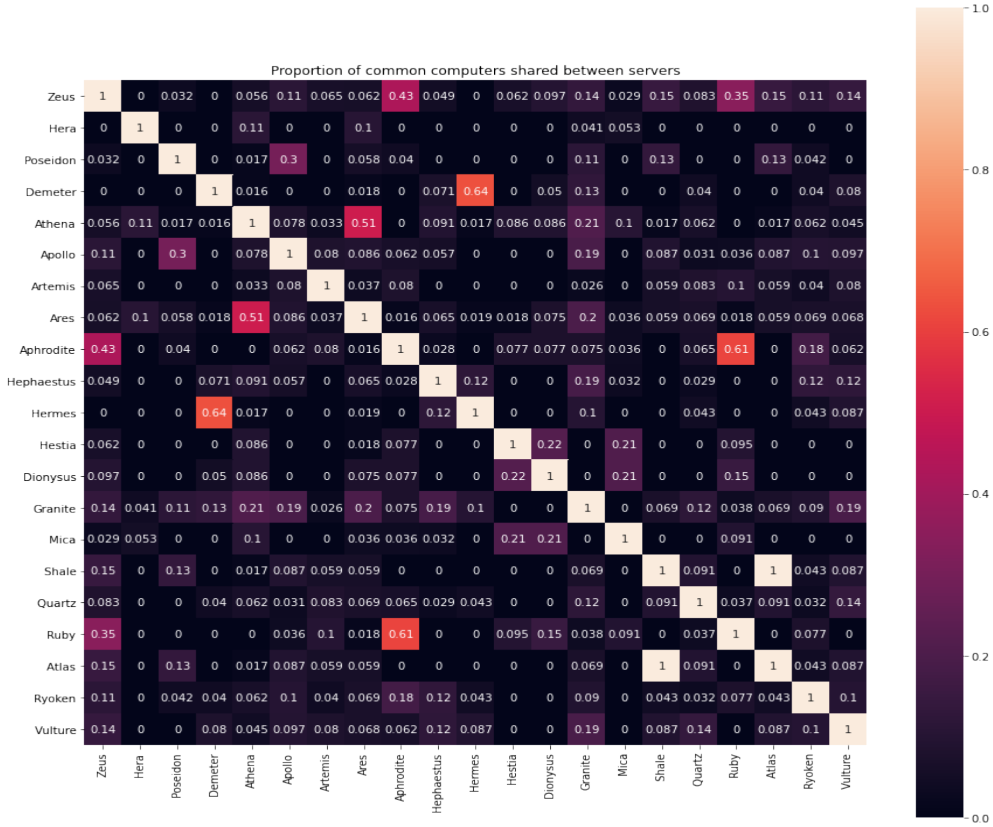


Figure 2: Visualisation of the Server Connections

Question 7 - Are there servers that share similar computers? What does the correlation look like?

The correlation map in Figure 3 shows the ratio between how many computers are shared between servers(intersection) over the union of the computers between the servers.



	Zeus	Hera	Poseidon	Demeter	Athena	Apollo	Artemis	Ares	Aphrodite	Hephaestus	Hermes	Hestia	Dionysus	Granite	Mica	Shale	Quartz	Ruby	Atlas	Ryoken	Vulture
Zeus	1	0	0.032	0	0.056	0.11	0.065	0.062	0.43	0.049	0	0.062	0.097	0.14	0.029	0.15	0.083	0.35	0.15	0.11	0.14
Hera	0	1	0	0	0.11	0	0	0.1	0	0	0	0	0	0.041	0.053	0	0	0	0	0	0
Poseidon	0.032	0	1	0	0.017	0.3	0	0.058	0.04	0	0	0	0	0.11	0	0.13	0	0	0.13	0.042	0
Demeter	0	0	0	1	0.016	0	0	0.018	0	0.071	0.64	0	0.05	0.13	0	0	0.04	0	0	0.04	0.08
Athena	0.056	0.11	0.017	0.016	1	0.078	0.033	0.51	0	0.091	0.017	0.086	0.086	0.21	0.1	0.017	0.062	0	0.017	0.062	0.045
Apollo	0.11	0	0.3	0	0.078	1	0.08	0.086	0.062	0.057	0	0	0	0.19	0	0.087	0.031	0.036	0.087	0.1	0.097
Artemis	0.065	0	0	0	0.033	0.08	1	0.037	0.08	0	0	0	0	0.026	0	0.059	0.083	0.1	0.059	0.04	0.08
Ares	0.062	0.1	0.058	0.018	0.51	0.086	0.037	1	0.016	0.065	0.019	0.018	0.075	0.2	0.036	0.059	0.069	0.018	0.059	0.069	0.068
Aphrodite	0.43	0	0.04	0	0	0.062	0.08	0.016	1	0.028	0	0.077	0.077	0.075	0.036	0	0.065	0.61	0	0.18	0.062
Hephaestus	0.049	0	0	0.071	0.091	0.057	0	0.065	0.028	1	0.12	0	0	0.19	0.032	0	0.029	0	0	0.12	0.12
Hermes	0	0	0	0.64	0.017	0	0	0.019	0	0.12	1	0	0	0.1	0	0	0.043	0	0	0.043	0.087
Hestia	0.062	0	0	0	0.086	0	0	0.018	0.077	0	0	1	0.22	0	0.21	0	0	0.095	0	0	0
Dionysus	0.097	0	0	0.05	0.086	0	0	0.075	0.077	0	0	0.22	1	0	0.21	0	0	0.15	0	0	0
Granite	0.14	0.041	0.11	0.13	0.21	0.19	0.026	0.2	0.075	0.19	0.1	0	0	1	0	0.069	0.12	0.038	0.069	0.09	0.19
Mica	0.029	0.053	0	0	0.1	0	0	0.036	0.036	0.032	0	0.21	0.21	0	1	0	0	0.091	0	0	0
Shale	0.15	0	0.13	0	0.017	0.087	0.059	0.059	0	0	0	0	0	0.069	0	1	0.091	0	1	0.043	0.087
Quartz	0.083	0	0	0.04	0.062	0.031	0.083	0.069	0.065	0.029	0.043	0	0	0.12	0	0.091	1	0.037	0.091	0.032	0.14
Ruby	0.35	0	0	0	0	0.036	0.1	0.018	0.61	0	0	0.095	0.15	0.038	0.091	0	0.037	1	0	0.077	0
Atlas	0.15	0	0.13	0	0.017	0.087	0.059	0.059	0	0	0	0	0	0.069	0	1	0.091	0	1	0.043	0.087
Ryoken	0.11	0	0.042	0.04	0.062	0.1	0.04	0.069	0.18	0.12	0.043	0	0	0.09	0	0.043	0.032	0.077	0.043	1	0.1
Vulture	0.14	0	0	0.08	0.045	0.097	0.08	0.068	0.062	0.12	0.087	0	0	0.19	0	0.087	0.14	0	0.087	0.1	1

Figure 3: Visualisation of the protein in the salted solution

It is interesting to see that "Atlas" and "Shale" share exactly the same computers, which means one acts as a duplicate server. From figures 1 and 2 we know that both ”Atlas” and ”Shale” have the least amount of computers in the network. Similarly, the second highest ranking score goes to ”Hermes” and ”Demeter” (8 and 10 computers respectively) with 7 computers in common. This suggest that a higher percentage occurs if the union of the computers between servers is relatively low. The only exception would be Ares and Athena with 50% of computers in common. Overall, the heatmap is quite ”dark”, suggesting that the servers do not share that many computers in common.

Link Analysis

We used link analysis to get answers to our questions about the relationship between the computers and servers. We calculated the different centralities such as degree, betweenness, closeness, page rank for our link analysis. At the end we then compared the results using a correlation chart to see which centralities have the same behaviour.

Question 1 - What does the data look like graphically?

In the first attempt to represent the data, we decided to build a graph that includes both servers and computers. In Figure 4, the servers are represented in cyan while the computers are the small circles shown in magenta.

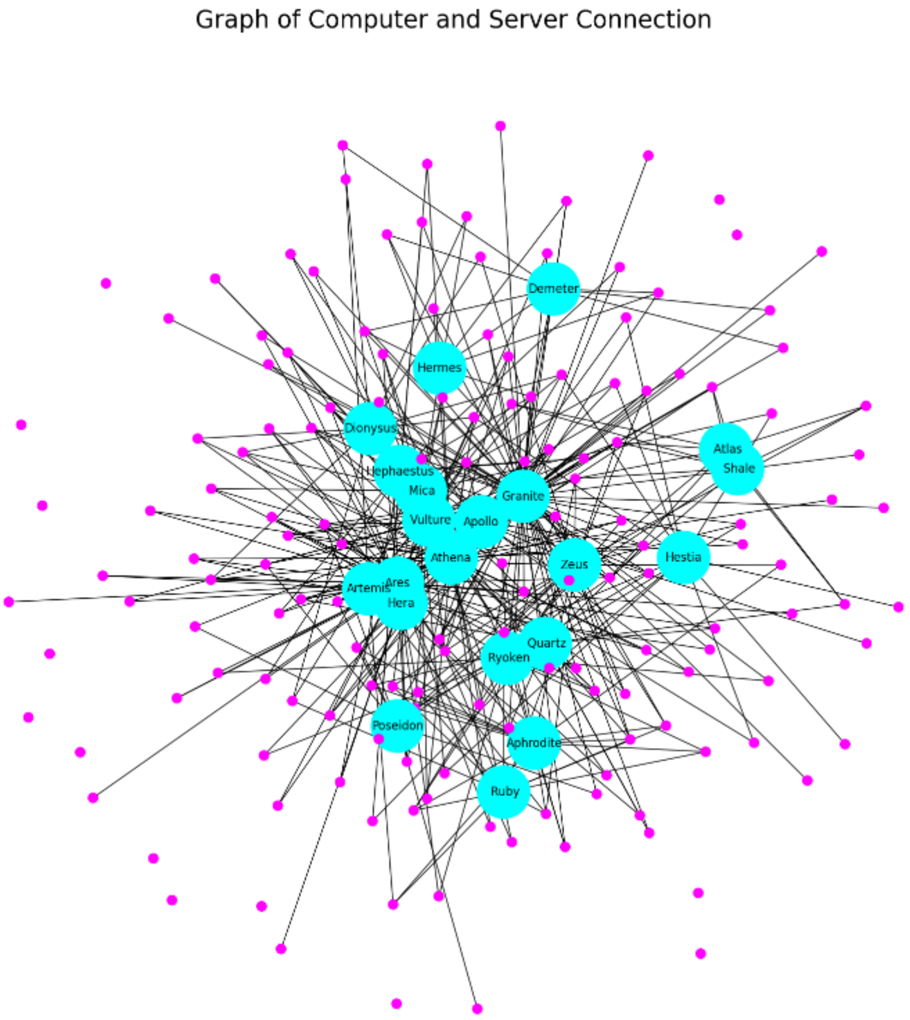


Figure 4: Graph showing the network between computers and servers

The result is messy and so no useful information can be obtained, other than, there are some computers that are isolates and computers that are only connected to one server.

Question 2 - Is there a better way to graphically represent the data?

Instead of showing the computers, we can make a network based solely on the servers (since the number of computers can vary more frequently than the number of servers in an Autonomous System). Instead the number of computers can be represented as an edge attribute where the thickness of the edge depends on the computers 2 servers (nodes) have in common. As a node attribute, the size of the node was represented as the total number of computers a server has (i.e. the degree of the server in the from Figure 4).

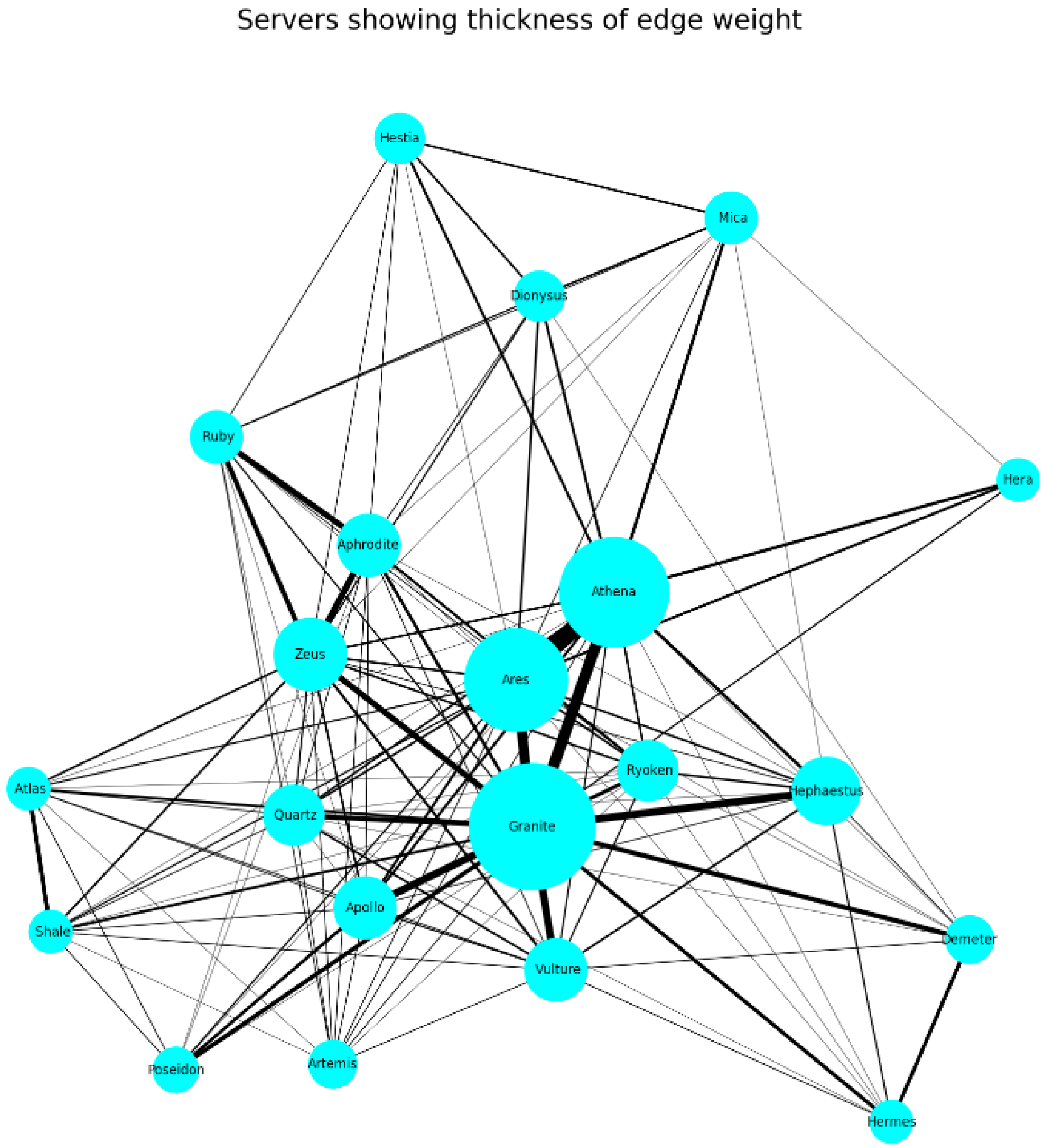


Figure 5: Graph showing the network between servers with computer connections as node attributes

This is a much better representation of the graph that clearly shows that the connection between servers can vary and the size of the server depends on how many computers it has.

Question 3 - Which one is the most important node in the graph?

The most basic measure of centrality is the degree centrality. According to this, in a network, a node with a larger number of edges is more important than a node with a smaller number of edges. This score is only local and strictly related to the node neighbourhood. Hence, while we can infer how a node with high degree centrality may affect its neighbours, we cannot say anything about more distant relationships.

In this specific graph the degree is computed by counting the number of connections between servers. Thus, the maximum degree a node can have in this graph is 20, as there are 21 servers. Figure 6 shows the graph with the node size represented as the degree centrality. (Since the proportion of size is difficult to notice, a color map was added to cluster similar sizes).

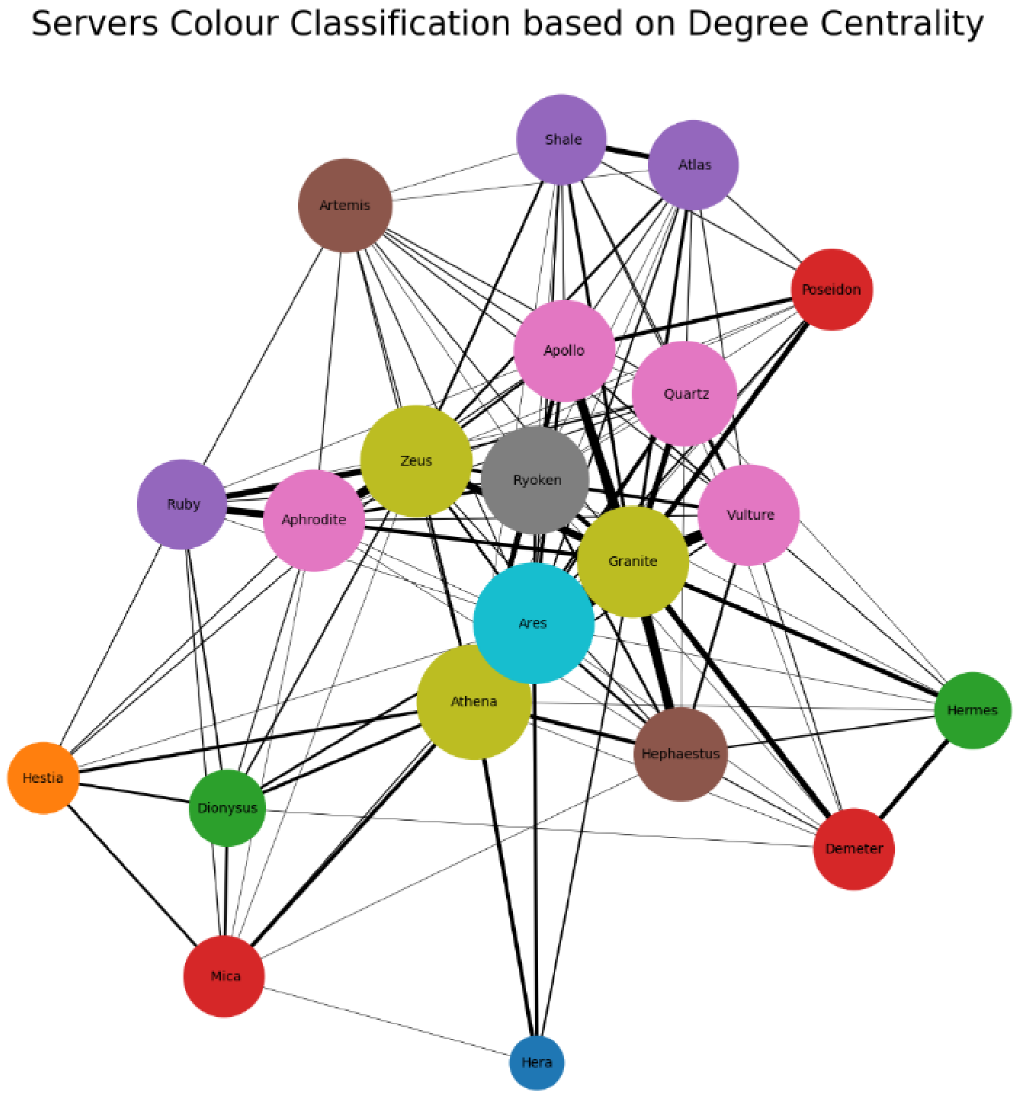


Figure 6: Graph based on Degree Centrality

”Ares” is the node with degree equal to 20, which mean its centrality is equal to one (1), classifying it as the most important server.

Question 4 - Is there another way to address the importance of a server?

Indeed there is. We used the PageRank centrality and added it as a node attribute. Differently from the degree centrality, the PageRank algorithm is useful because the importance of a node is directly related to the relevance of the the servers that link to it. In other words, the PageRank value represents the significance of a server in relation to all other servers. The size of the nodes in Figure 7 was computed through the PageRank values.

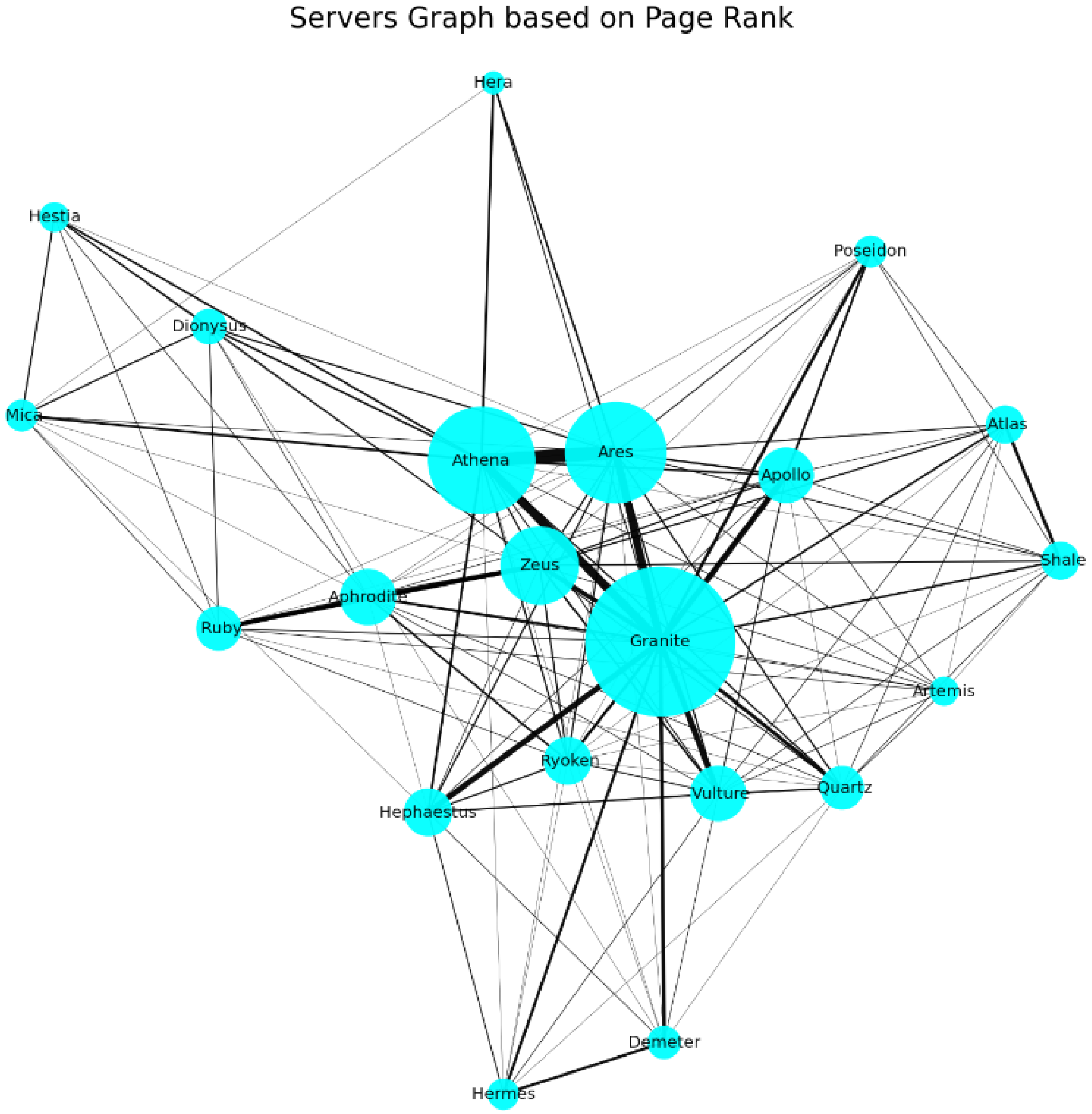


Figure 7: Graph based on PageRank Centrality

It is interesting to see that Granite is considered the most important node, as it takes the direct influence of Ares (most connected node) and all the other node connections it has. This graph is very similar in shape and node size to the one in Figure 5 (Question 2) which means that the PageRank graph not only represents the influence of other servers but the influence of the computers between them as well.

Question 5 - Are there any bridging positions in the network?

In intuitive terms, the betweenness centrality represents how often a node is found on the shortest path connecting two nodes. Therefore, a node with high betweenness centrality has high control over the networks because many connections pass through it. Nodes with high betweenness centrality are often considered as a ”bridge” in the network because they control the information being passed through.

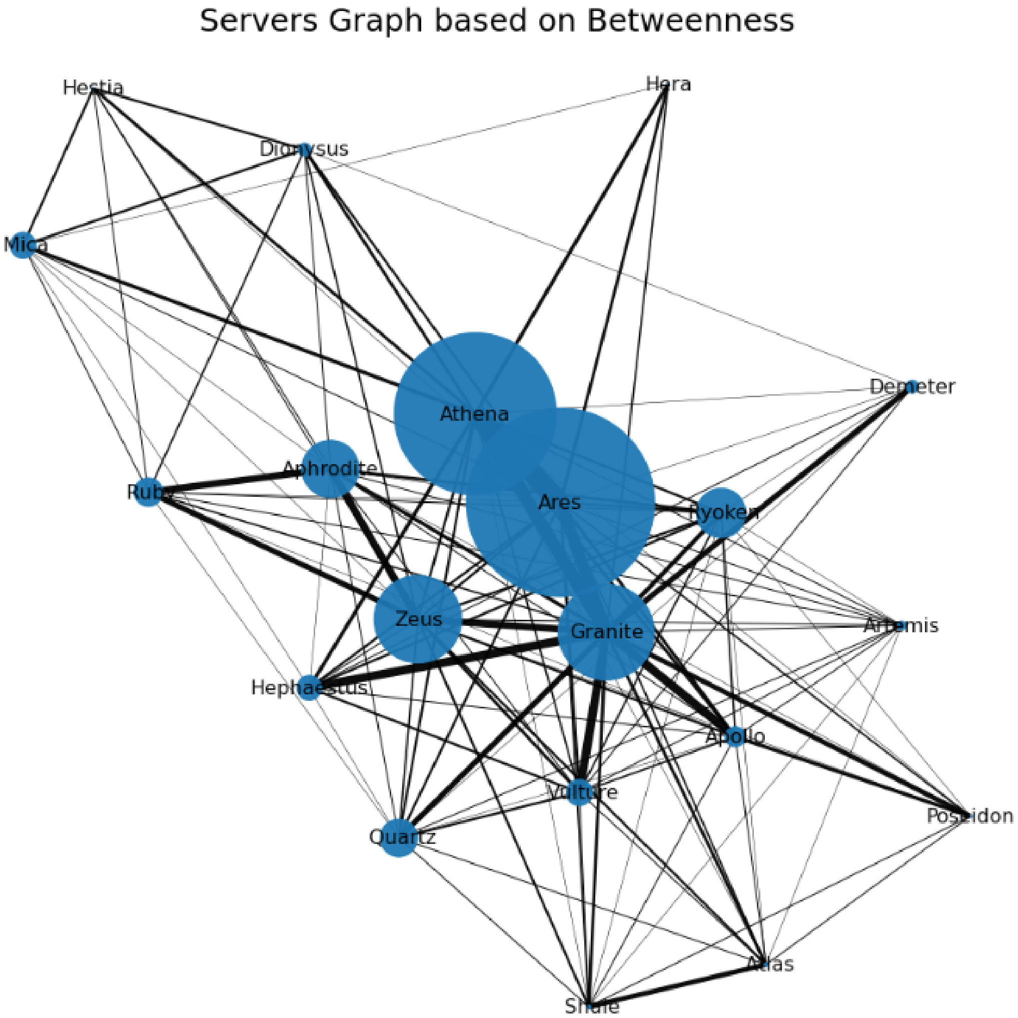


Figure 8: Graph based on Betweenness Centrality

The graph shown in Figure 8 is a little misleading because even though technically ”Ares” has the highest betweenness centrality (approx. 0.1), all the betweenness scores are relatively low (as shown in the histogram in Figure 10).

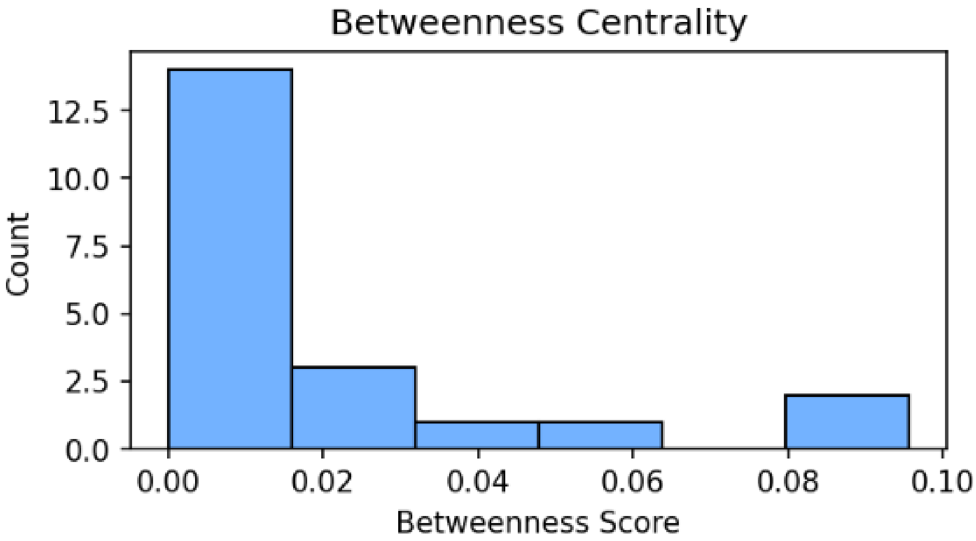


Figure 9: Histogram showing the distribution of the betweenness centrality among servers

This means that there are no bridging positions in the network and even if Ares is removed from it (e.g. the server failing), the network would still be very well connected.

Question 6 - How close is every server to another in the network?

The closeness centrality is defined as the sum of the length of the shortest paths between a node and any other node in a graph. A vertex with high closeness centrality is considered close to all other vertexes in the network.

The overall values of the closeness centrality was relatively high (higher than 0.6) with "Ares" having a value of 1. This was expected since "Ares" is fully connected in the network. Figure 10 shows the resulting graph with the closeness centrality as node attribute. (Similarly to Figure 6, since the proportion of size is difficult to notice, a color map was added to cluster similar sizes).

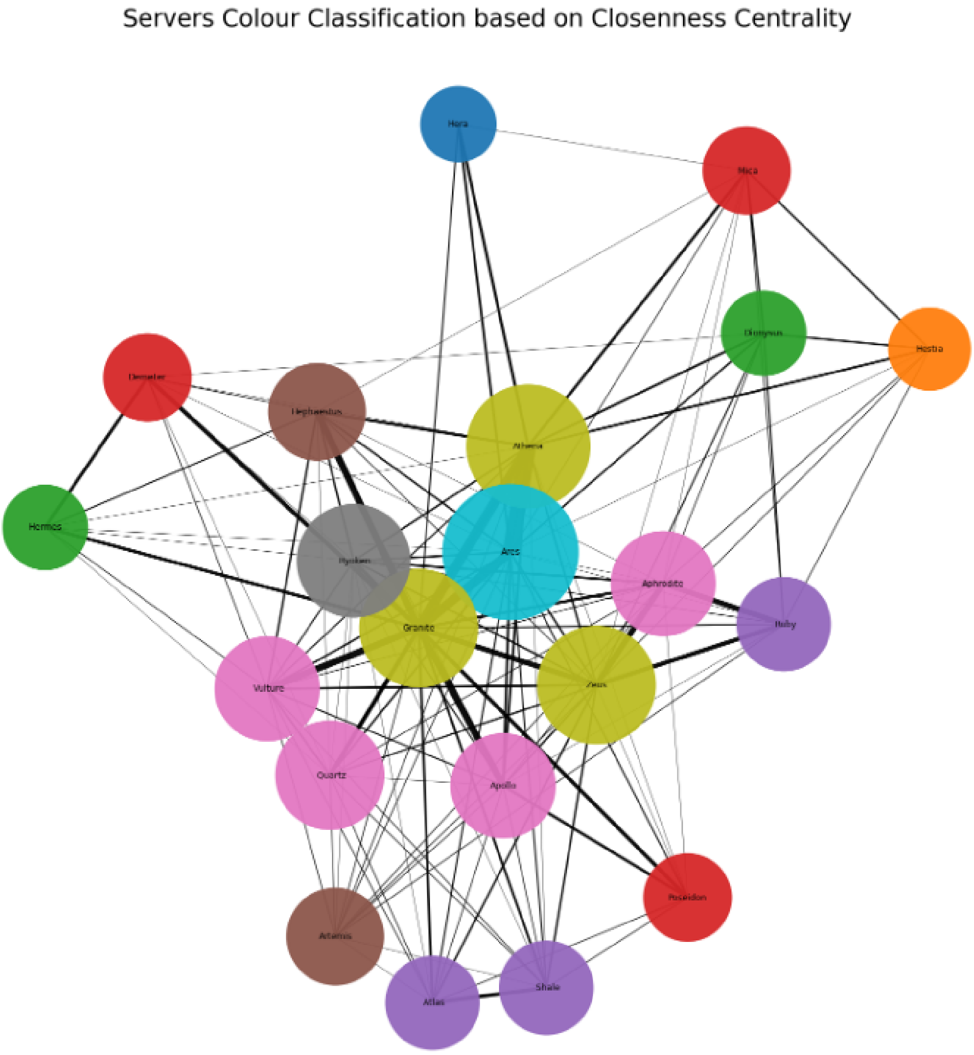


Figure 10: Closeness Centrality

Question 7 - What is the correlation between the centrality scores?

Figure 11 shows the correlation between all the centralities. From the image, we see that degree centrality is less correlated to page rank than the other centralities, but is highly correlated to closeness centrality. This means that the higher the page rank score, the more closer the nodes are too other nodes, showing that because the node is important according to its page rank score, it would need to be well connected to other nodes. Another correlation that is high is betweenness and closeness.

We noticed that page rank has the lowest correlation with the others, compared to the other correlations. This means that the popular nodes according to page rank might not be the highest nodes on the other centrality list.

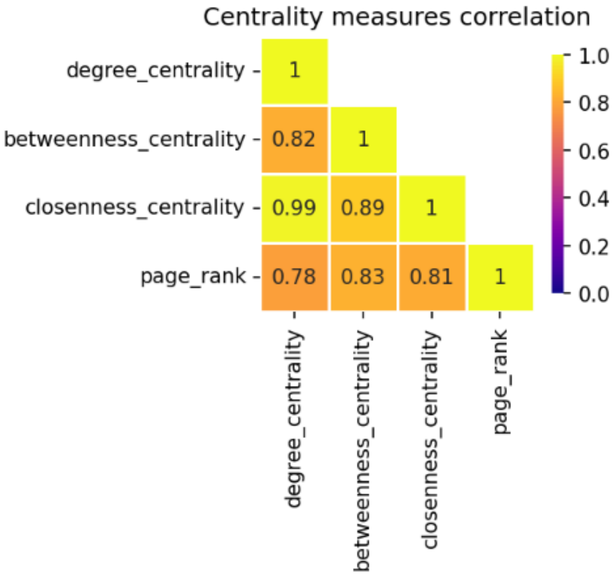


Figure 11: Comparison between Centralities

Conclusion

In conclusion, link analysis was very useful in helping us to learn more and visualise the relationship in our computer network. We can conclude from our results shown in the link analyses section that ”Granite” had the highest PageRank score and ”Ares” had the highest score with betweenness, degree and closeness score. The most important servers in our network is ”Granite” and ”Ares”.

The analysis and conclusion drawn from this report could prove useful if this network of Autonomous systems wants to be optimised. For example, adding safeguards to make sure that ”Granite” and ”Ares” do not fail or if more computers enter the network consider reallocating them to less connected servers.

Disclaimer

This report is based on the ’Autonomous_Systems.ipynb’ Jupyter notebook. Not all the results, graphs and plot are presented in the report due to the page limit, however detailed analysis and answers to various research questions are reflected here.

Therefore, the report is meant as a summary and analysis on the most important conclusions. Please refer to the ’Autonomous_Systems.ipynb’ notebook if inspection of other results is necessary.