Datathon 3

Sushranth Hebbar

Introduction :- Use tabular datasets published by the World Health Organization to create new networks and visualize network communities using tools like gephi and networkx.

Methods :- The approach to this task can be divided into 2 phases. The first phase is data preprocessing and the second is network visualisation. The first step consists of creating an adjacency matrix from the given data and generating a graph object after passing the matrix to networkx. The second step focuses on using gephi on the graph object to generate network visuals.

Preprocessing :- The covid_19_data.csv was used to generate the visuals. This is because this file contains the least amount of missing values and has time series data on the number of active cases and deaths country wise between January 22 to September 23. First, the dataset was loaded into a pandas dataframe. Then it was grouped based on the observed date. Then the confirmed cases for each province was summed up for every country for a particular time stamp was computed and stored in a dictionary. Then the square root transformation was applied on these values to ensure the stability of these values. Finally, they were converted into a dataframe and then the values were normalized.

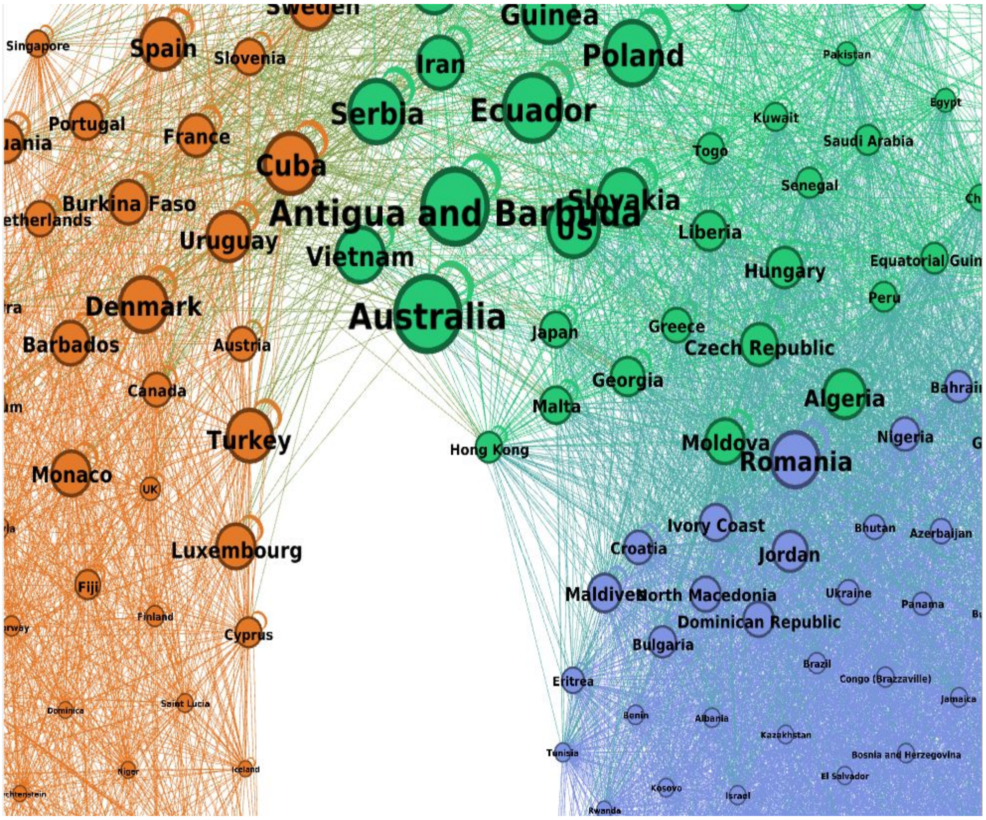
Then the correlation matrix was computed for the normalized value. The correlation matrix now contains values between -1 and 1 which indicates the extent of correlation between two countries. Now these values had to be filtered. The filter value was set to 0.95. That way, only countries with a high correlation value will have an edge between them. Now an adjacency matrix was computed where two countries have an edge between them only if their correlation value is at least 0.95. Otherwise, there will be no edge. Finally, this matrix is passed as an input to networkx and a graph object is returned. Then it was saved as a .gexf file. The time slice window used here was 10 days. So, a network was generated for a contiguous non overlapping window of 10 days. For the sake of inference, the visuals shown here a gap of 30 days between each other.

Network Visuals :-

The Fruchterman Reingold algorithm was applied on the graph. This algorithm is force based. So, nodes with large degree end up at the center of the graph and nodes with a low degree end up near the corners.After the graph stabilized, the betweenness centrality was computed for every node. Then the size of the node was determined based on the above centrality. This measures how often a node ends up on the shortest path between two nodes. So, it can determine the influence a node can have on a network. Next, the modularity was computed to observe communities and then the graph was colored accordingly. The labels were then applied on each node to identify the countries.These labels were scaled based on the size of the nodes. If the nodes correspond to countries, then the edges correspond to the similarity of the two adjacent nodes.

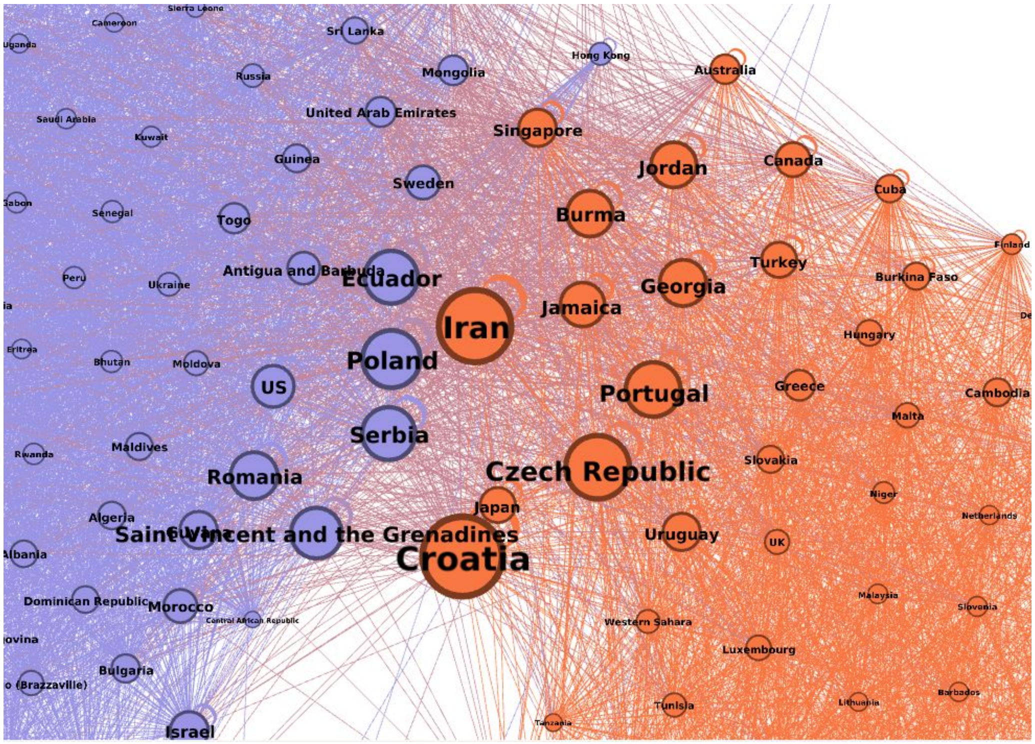
Inference :- All the inferences drawn below were derived from looking at the graph, wikipedia and raw data entries in the dataset. For more details look at reference.

Consider the time stamp January 22 - February 02.



During this point in time, the majority of the countries were free of the virus. But countries like Hong Kong, US, Australia, Japan and Vietnam were the first affected which makes them dissimilar with respect to other countries. So, they should have a smaller degree and based on how the FR algorithm works, they will get pushed towards the edge of their communities. Notice how the above mentioned countries are towards the edge of their green community. They tend to have a high centrality as they connect with nodes of other communities. The center of the communities have countries which are free of the virus.

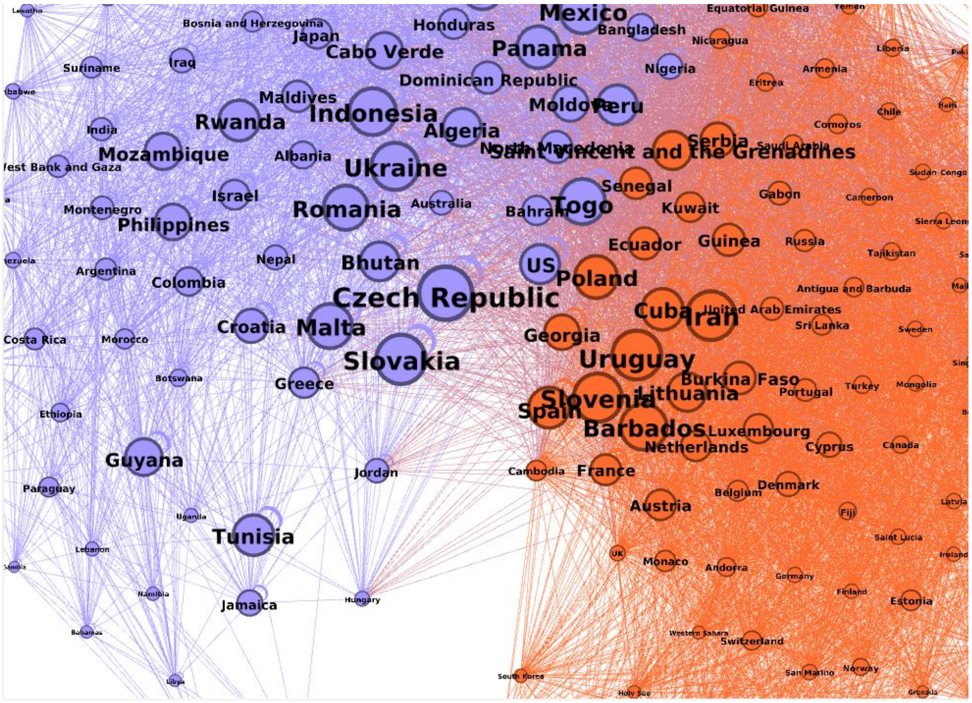
Consider February-22 to March-03



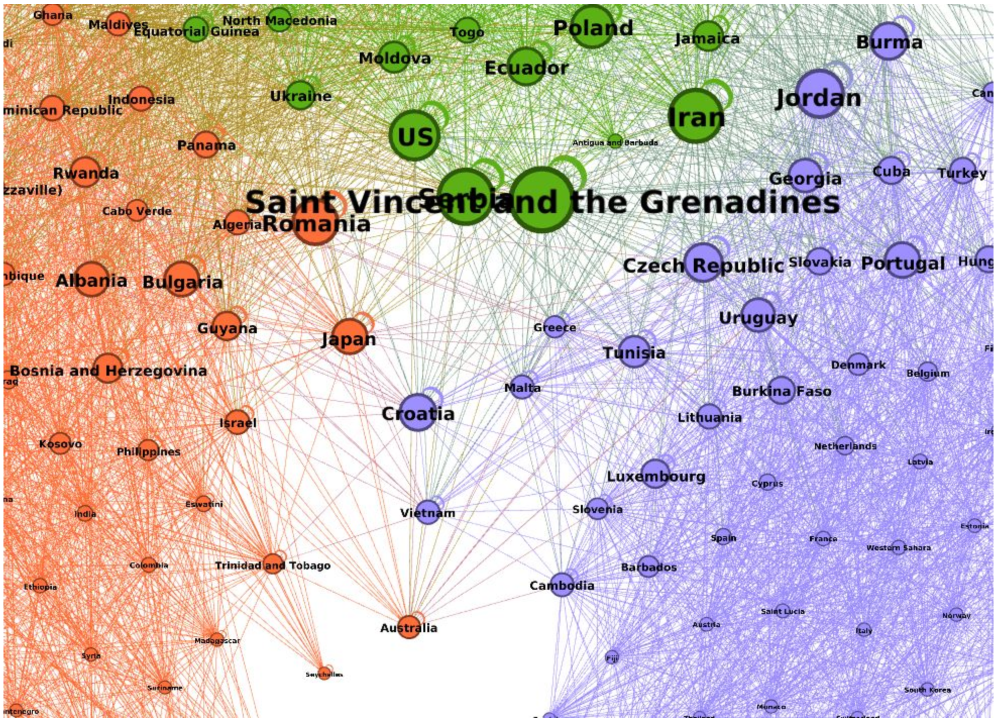
The configuration is different now. New countries like Croatia, Iran, Poland, Serbia, etc which were not near the edges are now here. More countries are getting infected by the virus. Countries like the US and Australia which were already affected have receded away from the edge of their communities but are still far away from the center. For instance Australia and Hong Kong,which is at the top right in the above image, is now at the opposite edge of it’s community. But there still exists a majority of countries which are not affected and which lie towards the center of their communities.

Consider March-22 to April-02

From the below image, it is clear that there are more nodes towards the edge of the communities. New countries like Slovenia, Barbados, Lithuania have appeared. This means that the infection is spreading to more countries and that the transmissions are accelerating in a few countries. Notice how Australia which was at top right in the previous image has shifted position again and is now back to where it was in the first image. The only difference is that it’s size is smaller. This means that the no of newly infected countries is higher than the countries where there is an increase in transmissions.

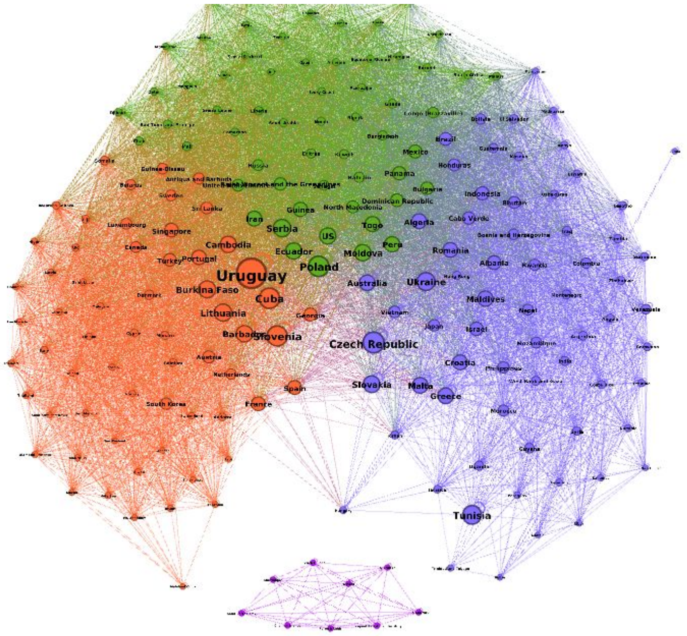


Consider April-22 to May-02



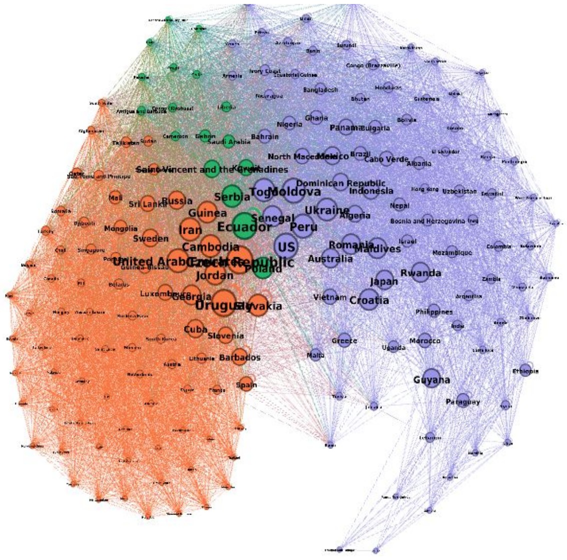
Notice how there are less nodes near the edge of the community compared to the previous image. This does not mean that the infections have reduced. Different communities are showing different trends. So, the countries towards the edges are definitely above stage 1 of transmission. As you move towards the center of each community, you will find countries which are still in stage 1. Another observation could be countries which are above stage 1 could be slowing down the virus in the form of curfews and a nation wide quarantine. So, the majority of countries are still in stage 1 and the few which are above this stage have started imposing restrictions. Look at how the position of countries like Australia and Japan which were the first to be affected have gradually changed in each image.

May-22 to Jun-02:

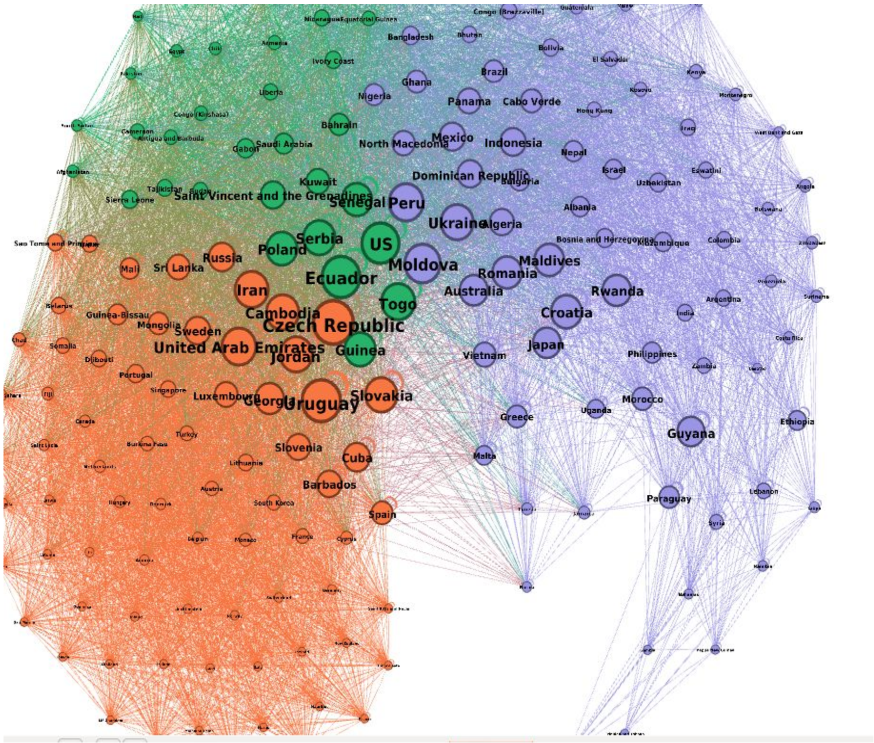


Notice the small cluster near the bottom. This indicates a small group of countries where the number of cases is not in the thousands. The rest of the countries have cases in the tens and hundreds of thousands.

June-22 to July-02:
The small cluster from the previous image has vanished. So, every country has at least a thousand confirmed cases. If you look at the center of the image, there are many nodes visible towards the edge of their respective communities which indicate a steady increase in the number of cases in these countries.

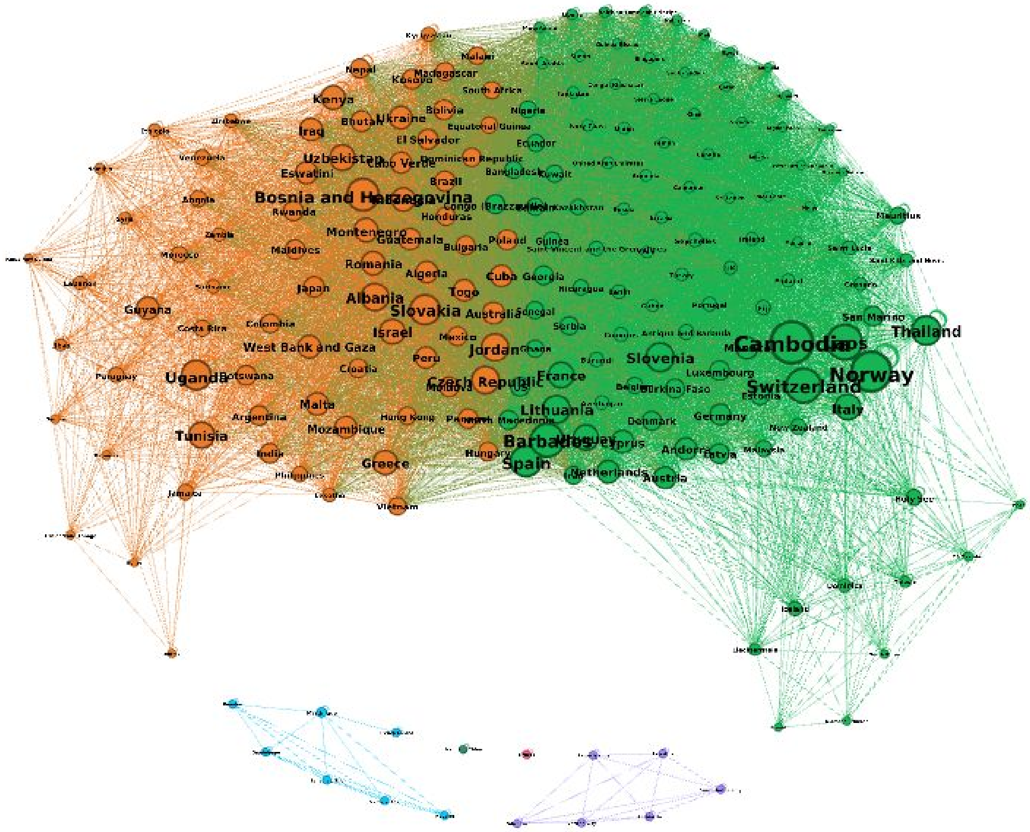


July-22 to Aug-02:



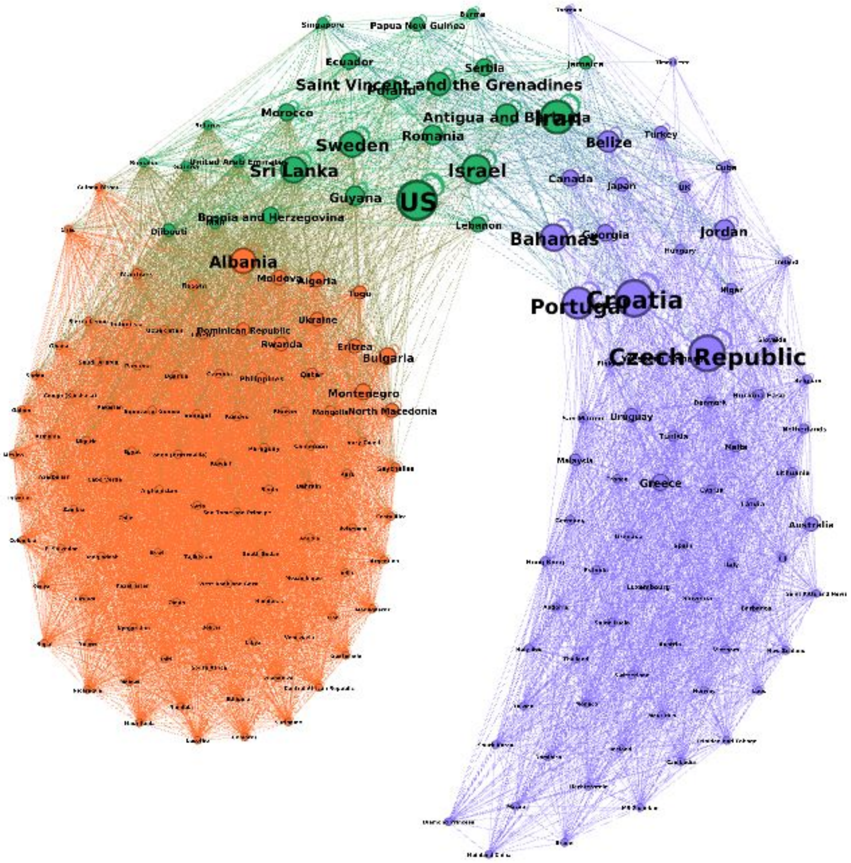
There is not any major difference between this and the previous image. (Notice how the position of Australia changes throughout the images)Some countries like Australia are suffering from a second wave of coronavirus while others like the US are struggling to contain the first wave.

August-22 to September-02



There exists new communities now near the bottom. The purple one's rid of the virus. The blue has managed to control the spread of the virus and has reduced the number of confirmed cases to a few hundred. More countries have started to bring the number of cases under control. There are more nodes towards the edges of the communities. This does not necessarily mean increased infection. It means there is a reversal in trend and the ones at the edges are controlling the virus and the ones at the center of the community are struggling to control it.

September 12- September 23 :-
Lots of countries have managed to bring the confirmed cases under control. Compared to the previous image, there are less nodes near the edges. Very few countries remain where the cases are actively rising. As you can see, the US is still struggling to contain the virus.



References :-

https://gephi.org/tutorials/gephi-tutorial-layouts.pdf
https://gephi.org/tutorials/gephi-tutorial-quick_start.pdf
https://networkx.github.io/documentation/stable/tutorial.html
https://en.wikipedia.org/wiki/COVID-19_pandemic_in_{country}

The last link was used to look at the statistics of few countries and confirm whether the inferences drawn were true or not.