Whitireia
NEW ZEALAND

WelTec

Te Pūkenga

DS6501 Social Data Analytics

Trimester 1

Assignment 2

Tutor: Abubakar Siddique

Due Date: 2 June 2024

Student: Adya Sinha

Student ID: 2230449

Word Count: 3422 words

Submission Date: 2 June 2024

Table Of Contents

Executive Summary

This report gives a comprehensive analysis of the dataset about the street gang network using the social data analytics techniques. The given data set has 2 files- *StreetGangLinks* and *StreetGangNodes* which were first imported and then analysed in RStudio with the *igraph* package.

The analysis involves us to create and visualise networks, calculate centrality measures, and identify ethnic communities within the different networks. The key findings are: -

- **Identification of key gang members**: The analysis identifies primary gang members who play a key role in the networks based on centrality measures. The top 3 gang members by degree centrality are nodes- 1,7,12; by betweenness centrality are- 1,12, 7; and by closeness centrality are- 12, 28, 25.
- **Ethnic Community Analysis**: Ranking the prison members by their birthplace revealed that the Caribbean and West African gang members have the most connections among all the other countries.
- **Authority scores and community detection**: Calculation of authority scores were done to determine the influence of gang members, and communities. The communities were detected to identify distinct groups within the network where it was concluded that the UK and Caribbean were the largest communities.
- **Network Visualisation**: The network visualisation revealed important clusters and interactions among gang members of the same and different ethnicities, particularly between the UK, West African, Caribbean and East African communities.

Additionally, the analysis indicates that while co-offending crimes occur among members of the same ethnicity, many significant interactions also take place across the ethnic boundaries.

Task One

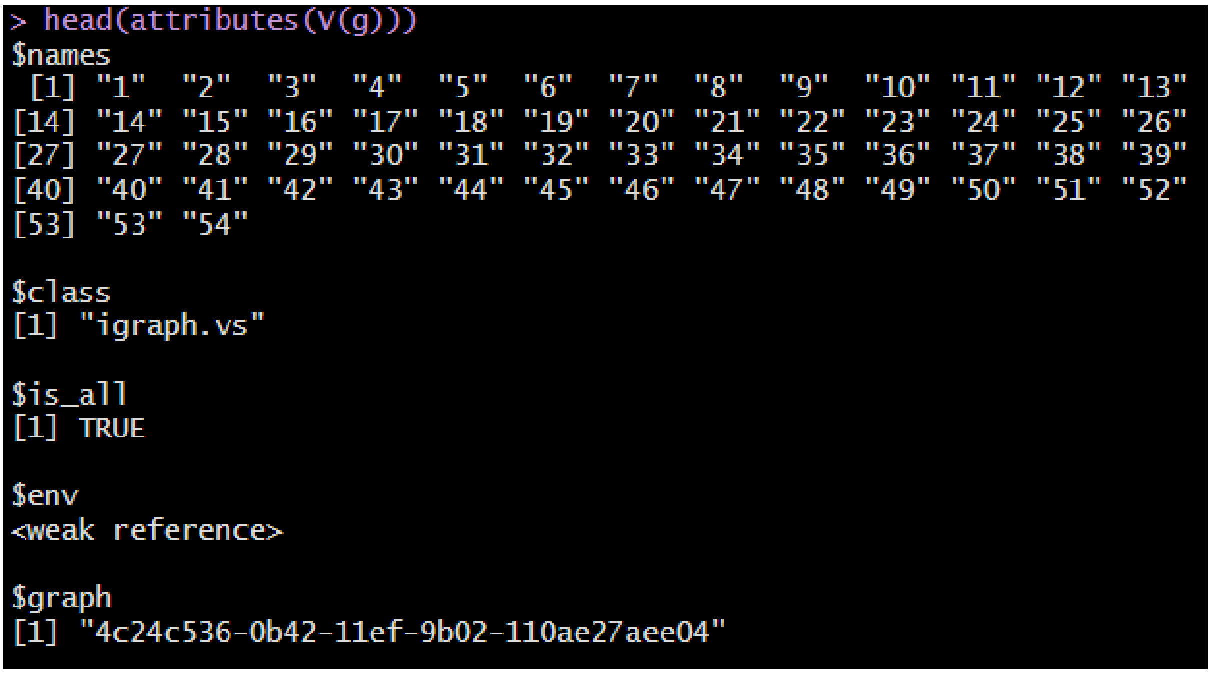
a.) The given dataset consists of two files: StreetGangLinks and StreetGangNodes. I used the *read.csv()* function in R to import these files after downloading it from Moodle.

b.) To create an igraph object, I used the *graph_from_data_frame()* function after downloading and loading the igraph package in R. Directed graphs are graphs where the edges have arrows showing the direction of the path. Undirected paths have bi-directional edges with no arrows.
Since the links should be treated as undirected, I set the directed parameter to FALSE.

c.) Networks are constituted of nodes and links where nodes represent points and links are represented as lines.

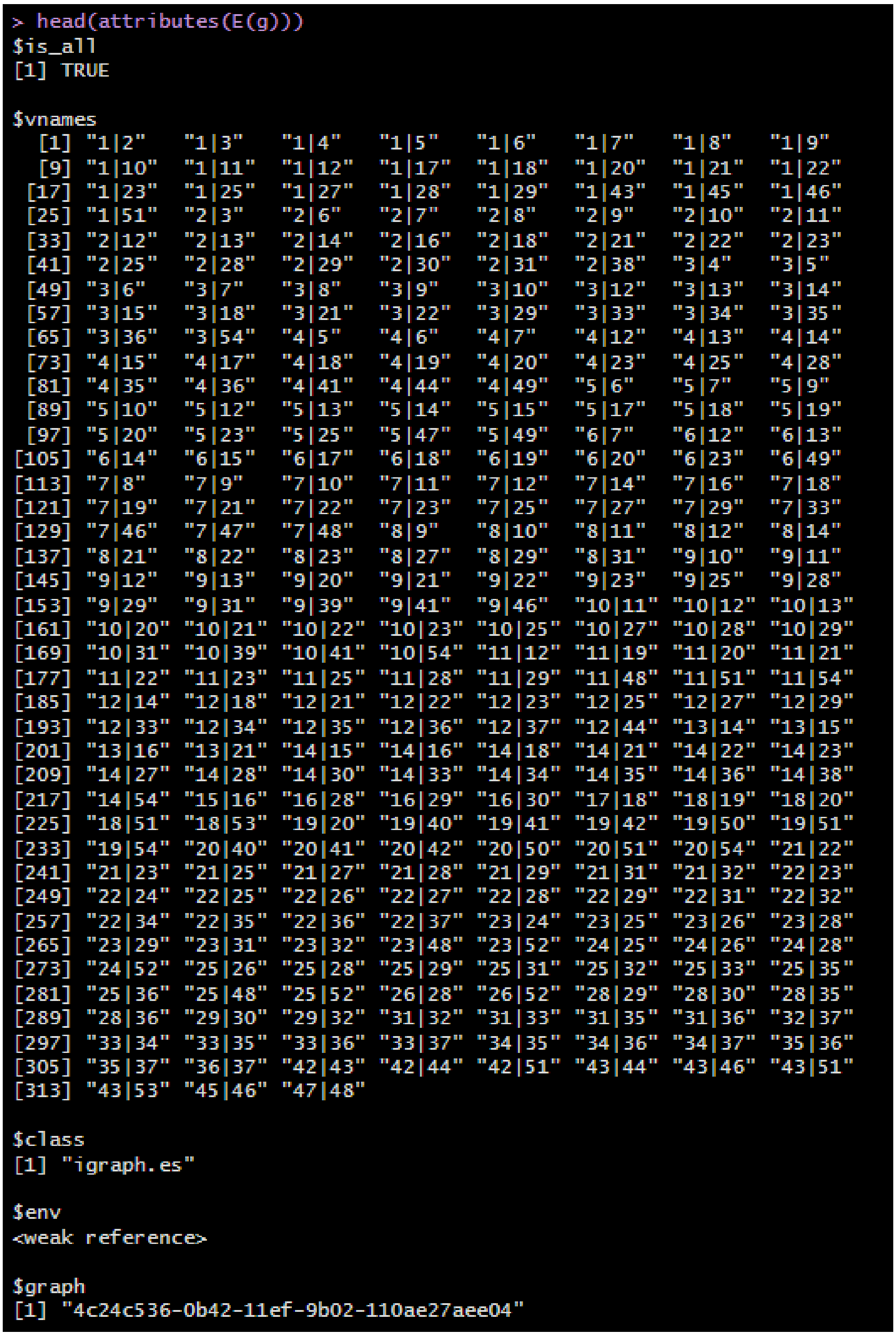
To inspect the attributes of the network, I used the *V()* and *E()* functions to access the nodes and links(edge list) respectively. Then using the *attributes()* function, we can retrieve the properties linked to those edges.This function will return a list of attribute names which can be used to access the values of the attributes.

As we can see in the figures below, the $names section shows us all the attributes of the nodes and links. As we can see, the *$is_all* section shows the boolean value “TRUE” which means that it is displaying all the attributes. The *$graph* section shows us a unique hash value related to the graph object.



```
> head(attributes(V(g)))
$names
 [1] "1"  "2"  "3"  "4"  "5"  "6"  "7"  "8"  "9"  "10" "11" "12" "13"
[14] "14" "15" "16" "17" "18" "19" "20" "21" "22" "23" "24" "25" "26"
[27] "27" "28" "29" "30" "31" "32" "33" "34" "35" "36" "37" "38" "39"
[40] "40" "41" "42" "43" "44" "45" "46" "47" "48" "49" "50" "51" "52"
[53] "53" "54"

$class
[1] "igraph.vs"

$is_all
[1] TRUE

$env
<weak reference>

$graph
[1] "4c24c536-0b42-11ef-9b02-110ae27aee04"
```

Figure 1 shows the node attributes



```
> head(attributes(E(g)))
$is_all
[1] TRUE

$vnames
  [1] "1|2"   "1|3"   "1|4"   "1|5"   "1|6"   "1|7"   "1|8"   "1|9"  
  [9] "1|10"  "1|11"  "1|12"  "1|17"  "1|18"  "1|20"  "1|21"  "1|22" 
 [17] "1|23"  "1|25"  "1|27"  "1|28"  "1|29"  "1|43"  "1|45"  "1|46" 
 [25] "1|51"  "2|3"   "2|6"   "2|7"   "2|8"   "2|9"   "2|10"  "2|11" 
 [33] "2|12"  "2|13"  "2|14"  "2|16"  "2|18"  "2|21"  "2|22"  "2|23" 
 [41] "2|25"  "2|28"  "2|29"  "2|30"  "2|31"  "2|38"  "3|4"   "3|5"  
 [49] "3|6"   "3|7"   "3|8"   "3|9"   "3|10"  "3|12"  "3|13"  "3|14" 
 [57] "3|15"  "3|18"  "3|21"  "3|22"  "3|29"  "3|33"  "3|34"  "3|35" 
 [65] "3|36"  "3|54"  "4|5"   "4|6"   "4|7"   "4|12"  "4|13"  "4|14" 
 [73] "4|15"  "4|17"  "4|18"  "4|19"  "4|20"  "4|23"  "4|25"  "4|28" 
 [81] "4|35"  "4|36"  "4|41"  "4|44"  "4|49"  "5|6"   "5|7"   "5|9"  
 [89] "5|10"  "5|12"  "5|13"  "5|14"  "5|15"  "5|17"  "5|18"  "5|19" 
 [97] "5|20"  "5|23"  "5|25"  "5|47"  "5|49"  "6|7"   "6|12"  "6|13" 
[105] "6|14"  "6|15"  "6|17"  "6|18"  "6|19"  "6|20"  "6|23"  "6|49" 
[113] "7|8"   "7|9"   "7|10"  "7|11"  "7|12"  "7|14"  "7|16"  "7|18" 
[121] "7|19"  "7|21"  "7|22"  "7|23"  "7|25"  "7|27"  "7|29"  "7|33" 
[129] "7|46"  "7|47"  "7|48"  "8|9"   "8|10"  "8|11"  "8|12"  "8|14" 
[137] "8|21"  "8|22"  "8|23"  "8|27"  "8|29"  "8|31"  "9|10"  "9|11" 
[145] "9|12"  "9|13"  "9|20"  "9|21"  "9|22"  "9|23"  "9|25"  "9|28" 
[153] "9|29"  "9|31"  "9|39"  "9|41"  "9|46"  "10|11" "10|12" "10|13"
[161] "10|20" "10|21" "10|22" "10|23" "10|25" "10|27" "10|28" "10|29"
[169] "10|31" "10|39" "10|41" "10|54" "11|12" "11|19" "11|20" "11|21"
[177] "11|22" "11|23" "11|25" "11|28" "11|29" "11|48" "11|51" "11|54"
[185] "12|14" "12|18" "12|21" "12|22" "12|23" "12|25" "12|27" "12|29"
[193] "12|33" "12|34" "12|35" "12|36" "12|37" "12|44" "13|14" "13|15"
[201] "13|16" "13|21" "14|15" "14|16" "14|18" "14|21" "14|22" "14|23"
[209] "14|27" "14|28" "14|30" "14|33" "14|34" "14|35" "14|36" "14|38"
[217] "14|54" "15|16" "16|28" "16|29" "16|30" "17|18" "18|19" "18|20"
[225] "18|51" "18|53" "19|20" "19|40" "19|41" "19|42" "19|50" "19|51"
[233] "19|54" "20|40" "20|41" "20|42" "20|50" "20|51" "20|54" "21|22"
[241] "21|23" "21|25" "21|27" "21|28" "21|29" "21|31" "21|32" "22|23"
[249] "22|24" "22|25" "22|26" "22|27" "22|28" "22|29" "22|31" "22|32"
[257] "22|34" "22|35" "22|36" "22|37" "23|24" "23|25" "23|26" "23|28"
[265] "23|29" "23|31" "23|32" "23|48" "23|52" "24|25" "24|26" "24|28"
[273] "24|52" "25|26" "25|28" "25|29" "25|31" "25|32" "25|33" "25|35"
[281] "25|36" "25|48" "25|52" "26|28" "26|52" "28|29" "28|30" "28|35"
[289] "28|36" "29|30" "29|32" "31|32" "31|33" "31|35" "31|36" "32|37"
[297] "33|34" "33|35" "33|36" "33|37" "34|35" "34|36" "34|37" "35|36"
[305] "35|37" "36|37" "42|43" "42|44" "42|51" "43|44" "43|46" "43|51"
[313] "43|53" "45|46" "47|48"

$class
[1] "igraph.es"

$env
<weak reference>

$graph
[1] "4c24c536-0b42-11ef-9b02-110ae27aee04"
```

Figure 2 shows the link attributes

d.) Using the *plot()*, we can create a plot of the network g with the default settings.

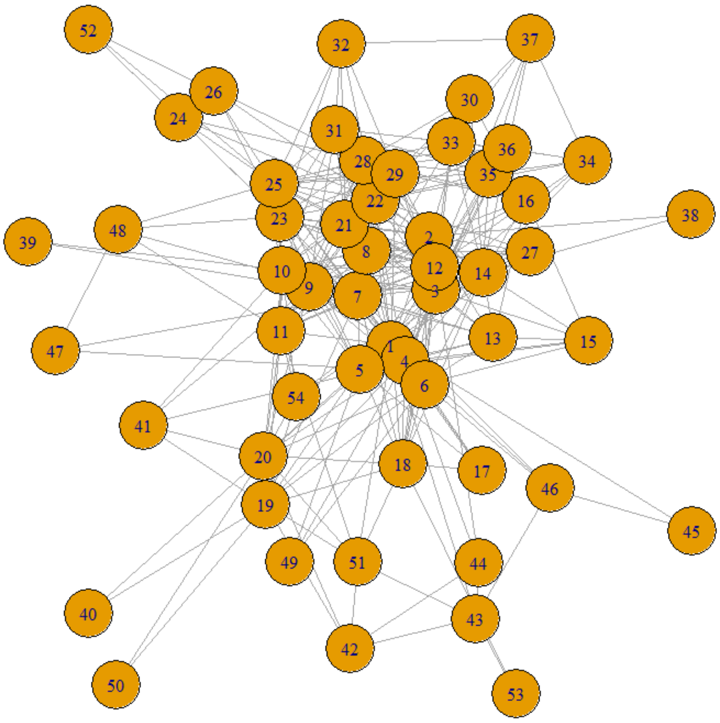


Figure 3 shows the default settings plot

When it comes to readability, there are several issues with this default plot which are as given below: -

- We can see that there are a lot of overlapping Nodes and Links since the network has many nodes and links. This makes it hard to distinguish between them.
- Since by default, all nodes and links are plotted with the same colour and size, it does not have significant colours and sizes. This makes it hard for the plot to convey any information about the attributes of the links and nodes like the age of the gang members, etc.
- The lack of labels makes it hard to identify which node corresponds to which gang member, or to know what the specific links represent.
- The default layout algorithm does not provide a good visualisation for this given network. As we can see, there are some nodes that are too close, while some are too far apart.

e.) Plotting a network again using the new attribute settings will create a plot g where the width of the links between nodes is set to the value of their weight attribute, and the size of each node is equal to the Age attribute divided by 2 which is shown below: -

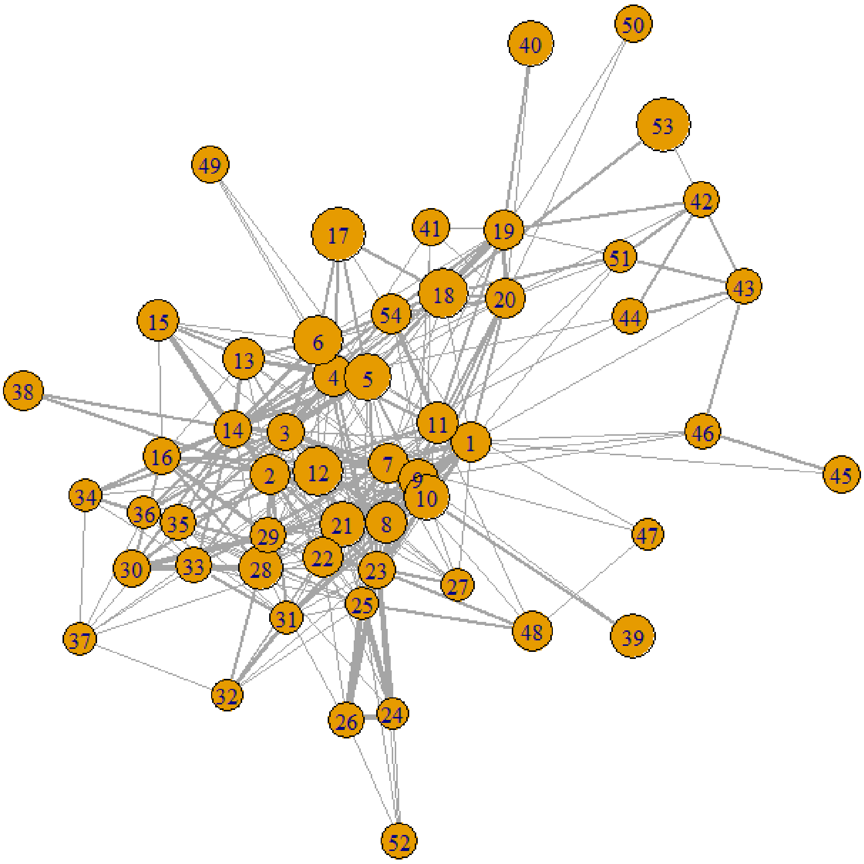


Figure 4 shows the default plot with some changes

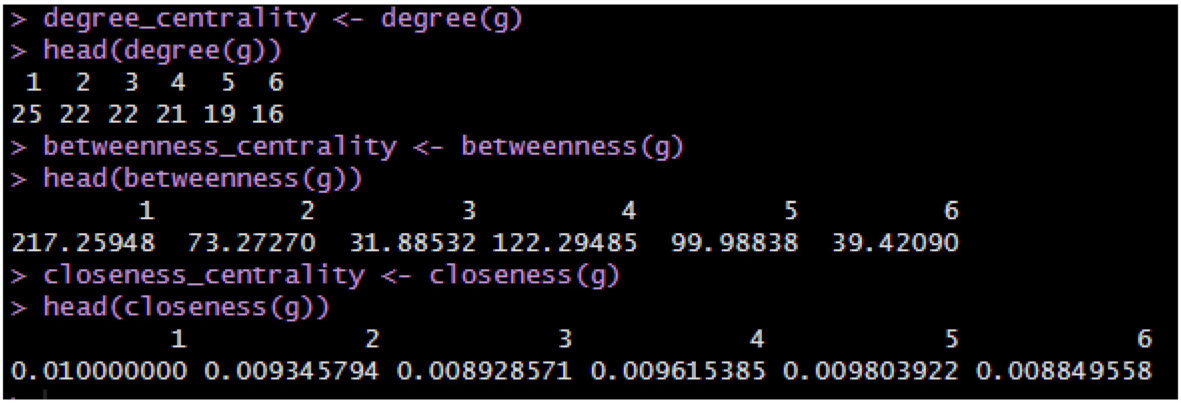
Although some issues might remain, such as the overlapping of nodes, lack of labels, and layout issues, the plot has been improved in the following ways: -

- The width of the links now represents the weight attribute, which makes it easier to identify the most significant interactions in the network. This could be interpreted as the strength or frequency of interaction between the nodes.
- The size of the nodes now represents the Age attribute which could help identify the older members of the gang, who might play a more central or influential role.

Task Two

a.) Centrality of a node measures the significance of the given node in the network. The degree centrality shows how connected a node is by counting the number of edges. The higher the degree, the more connections that gang member has. Betweenness centrality basically measures how frequently a node lies on the quickest path in between other pairs of nodes in the network. Closeness centrality measures how close the nodes are to each other in the network, based on the sum of the shortest path distances from that node to the other nodes.

To explore the measures of centrality within the network by calculating the degree, betweenness, and closeness measures of each node, we use the following functions as seen in *figure 5*: -

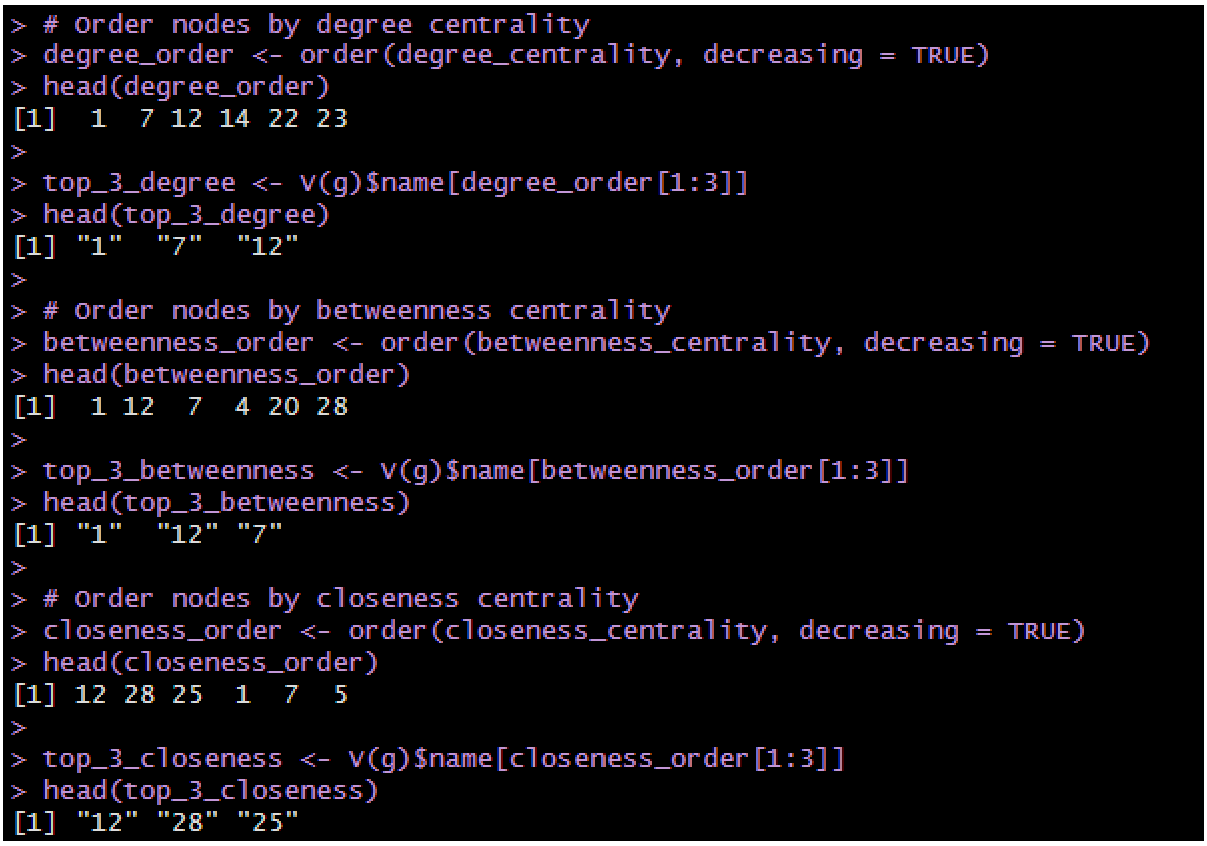


```
> degree_centrality <- degree(g)
> head(degree(g))
 1  2  3  4  5  6 
25 22 22 21 19 16 
> betweenness_centrality <- betweenness(g)
> head(betweenness(g))
        1         2         3         4         5         6 
217.25948  73.27270  31.88532 122.29485  99.98838  39.42090 
> closeness_centrality <- closeness(g)
> head(closeness(g))
          1           2           3           4           5           6 
0.010000000 0.009345794 0.008928571 0.009615385 0.009803922 0.008849558 
```

Figure 5 shows the different measures of centrality

b.) The top 3 nodes of the highest values for each centrality measure(degree, betweenness and closeness) are given below: -

- Nodes with high degree centrality are very well connected which means that the nodes **1,7, and 12** play a crucial role in communication and flow of information in the network.
- Nodes that have a high betweenness centrality act as bridges controlling the flow of information which means that nodes **1,12, and 7** control the flow of information between different parts of the given network.
- Nodes with a high closeness centrality quickly reaches other nodes which means that the nodes **12, 28, and 25** are possibly influential to spread information through the given network.



```
> # Order nodes by degree centrality
> degree_order <- order(degree_centrality, decreasing = TRUE)
> head(degree_order)
[1]  1  7 12 14 22 23
>
> top_3_degree <- V(g)$name[degree_order[1:3]]
> head(top_3_degree)
[1] "1"  "7"  "12"
>
> # Order nodes by betweenness centrality
> betweenness_order <- order(betweenness_centrality, decreasing = TRUE)
> head(betweenness_order)
[1]  1 12  7  4 20 28
>
> top_3_betweenness <- V(g)$name[betweenness_order[1:3]]
> head(top_3_betweenness)
[1] "1"  "12" "7"
>
> # Order nodes by closeness centrality
> closeness_order <- order(closeness_centrality, decreasing = TRUE)
> head(closeness_order)
[1] 12 28 25  1  7  5
>
> top_3_closeness <- V(g)$name[closeness_order[1:3]]
> head(top_3_closeness)
[1] "12" "28" "25"
```

Figure 6 shows the order nodes and top 3 nodes for each measure

Identifying these gang members may be of interest to the police because of the following reasons: -

- Degree Centrality: The gang members are connected to others and they could be key players in the gang activities. Nodes with high degree centrality are involved in a lot of interactions.
- Betweenness Centrality: The members could be important in the flow of information or resources within the gang. Nodes that have a high betweenness centrality could act as bridges between other nodes.
- Closeness Centrality: The gang members could have a lot of influence within the gang due to their communication abilities. Nodes(gang members) with high closeness centrality can reach all other nodes in the network quickly.

Task Three

a.) First we simplify the network by removing any nodes with a degree less than 15. Then we remove any edges whose weight attribute is less than 3 to then plot the network using the layout option of *layout_nicely* to get the graph shown below: -

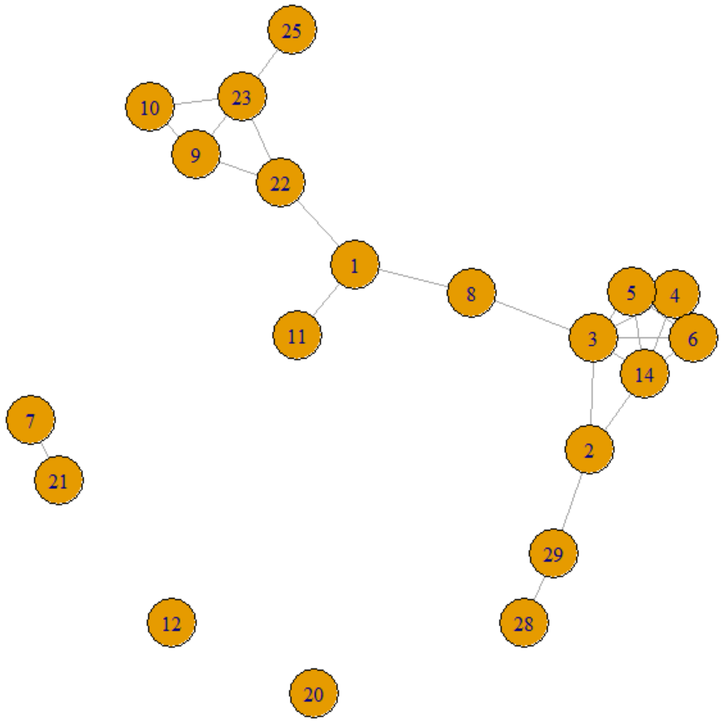


Figure 7

b.) The above network visualisation shows us the connection between different group members which represents the interactions within the gangs. As we can see, there are two major groupings of the nodes. The node labelled as ‘8’ is a bridge between these two main groupings, connecting them through its edges to the nodes ‘9’, ‘22’, and ‘1’in the smaller cluster and nodes ‘3’, ‘6’, and ‘28’ in the larger cluster.

Based on the analysis performed in Task 2, we can see that ‘12’ is the most influential gang member within the network as it is the most central position with the highest degree centrality. It has a lot of connections and is very influential within the gang.

Task Four

a.) First, I created a vector *ranking_colors* with colors that go with each ranking value and then assigned the color to each node's color attributes based on their ranking value. After plotting the network, I also added a legend which explains the colors used in the different node rankings. The red nodes have a ranking of 1, being the most important one followed by the green nodes with a ranking of 2 before the orange nodes which have a ranking of 3. The purple nodes have a ranking of 4 followed by the lowest ranking node of 5 which are the pink nodes.

The higher ranking nodes which are colored red and green are positioned centrally which means they have more direct connections to the other nodes(gang members). This indicates that the higher- assigned ranking gang members are involved in more co-offending activities than the other gang members.

The orange nodes are spread all across the network indicating that they have multiple connections.

The lower ranking nodes represented by the colors purple and pink are less important and are mostly located on the outer edges of the given network. This means they have less central role and have fewer co-offending activities with the other gang members and have a lower status in the gang hierarchy.

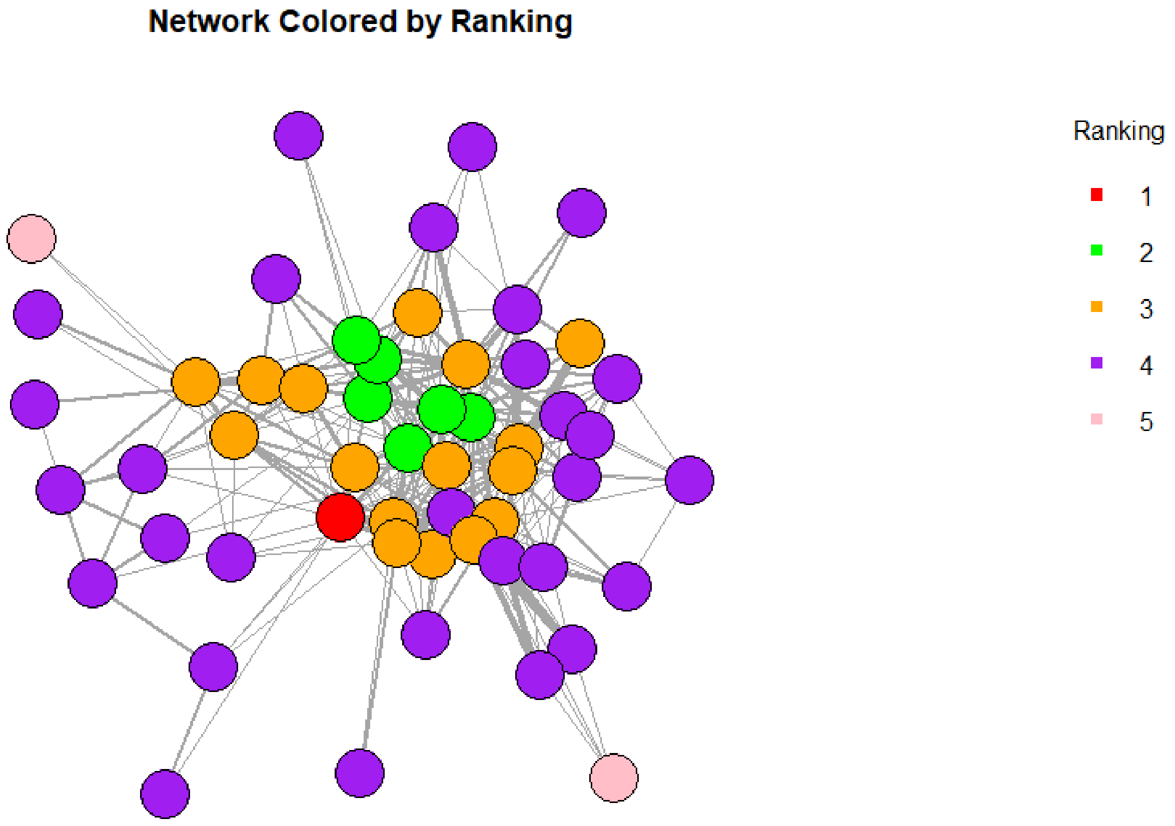


Figure 8 shows the networks colored by their assigned ranking value b.) The given simplified graph shows the groupings based on the birthplace with some ethnic clustering. The first step was to create a vector *birthplace_colors* with colors relating to the birthplace values. Then we assign the colors to each nodes based on their birthplace value. The next step was to create a subgraph with only gang members who have served prison time, then a legend was also added to the plot which helps us observe interactions between the gang members of varied ethnicities and if they interact with each other. The node colors are in four categories- brown for "1", cyan ranking "2", magenta for the third ranking and fourth ranking denoted by the color grey.

We can see that there is a concentration of brown(West Africa) and cyan (Caribbean) colors in the centre along with some scattered connections indicating that the gang members from these birthplaces have a lot of connections in the network.

Magenta (UK) and grey(East African) nodes are seen to appear less central which means that there are less connections in this network.

It can also be noted that there is some ethnic grouping visible as there seems to be some clustering as many different colored nodes are closer together. Many inter-ethnic interactions can also be seen as there are many connections of varied colors.

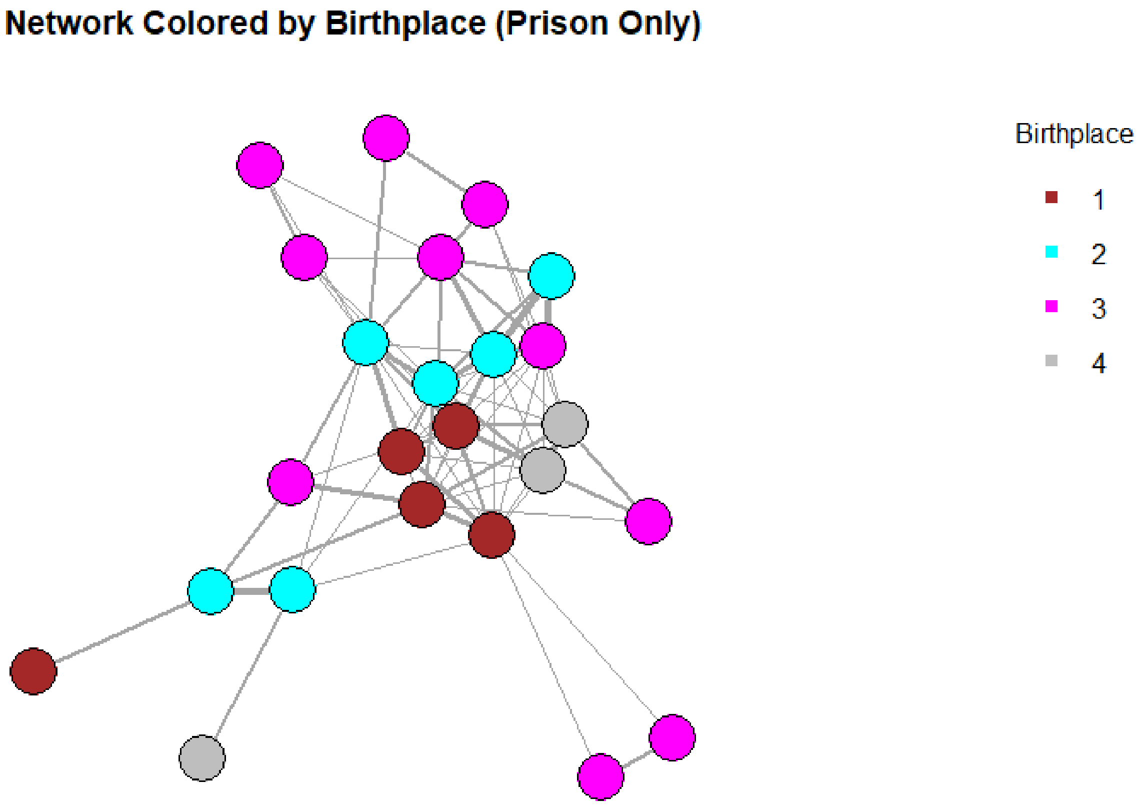


Figure 9 shows the network colored by birthplace c.) The first step is to create a subgraph which only includes the nodes (gang members) with a ranking of 3 or higher and then calculating the hub scores for the nodes using the *hub_scores* function. After creating the 2-panel plot, we plot the subgraph, setting the vertex.size parameter to 15 times the value of the hub scores.

In the first panel plot, each node represents a gang member and the node size is proportional to the member's hub score. The larger nodes like cyan and magenta show they have a higher hub score and are located centrally which means that those members have a central role in the network's connectivity with many connections to other members.

The second panel on the right shows the different communities in the network, where they are each highlighted in different colors. The communities were developed using the *cluster_optimal* function. We can observe some larger clusters link the yellow and magenta clusters while also some smaller clusters like the purple and green. As we can see, there are many smaller communities connected through the hub nodes making the network appear decentralised. The magenta and yellow are the two communities where gang members possess higher hub scores as compared to other communities in this network. The yellow community has the largest hub node(node 54) which tells us that it plays a primary role in connecting different parts of the network and has a high influence. The magenta community is smaller than the yellow community, but still pretty large and has some highly influential nodes as seen by its hub scores with the largest hub node being node 48.

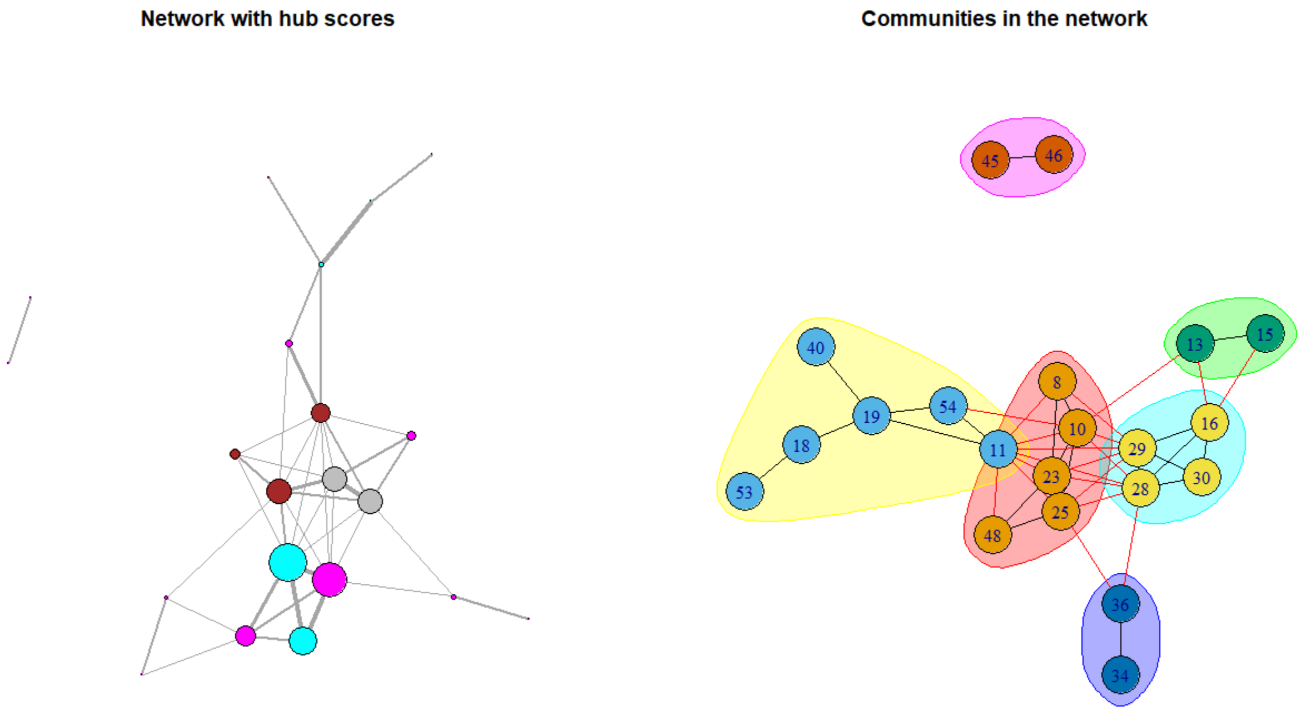


Figure 10 showing a 2-panel plot

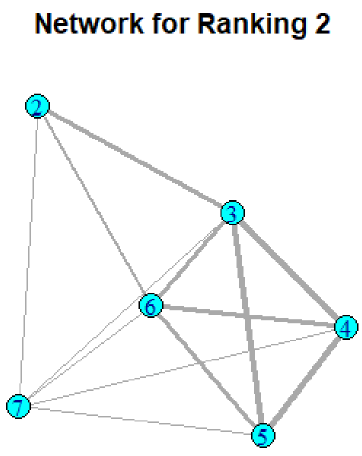
Task Five

First, we define a color palette for the graphs by using the colors_ranking function. To simplify the network from an earlier network, we first create the subgraphs for each ranking and then plot it by filtering the nodes by ranking.

- Ranking 1 (Red)
  Observation: I don't agree with the assigned ranking value as we can see, the network for ranking 1 is very sparse with no connections.
  Justification: Gang members with this ranking maybe less involved in any serious co-offending activities, thus fewer connections.

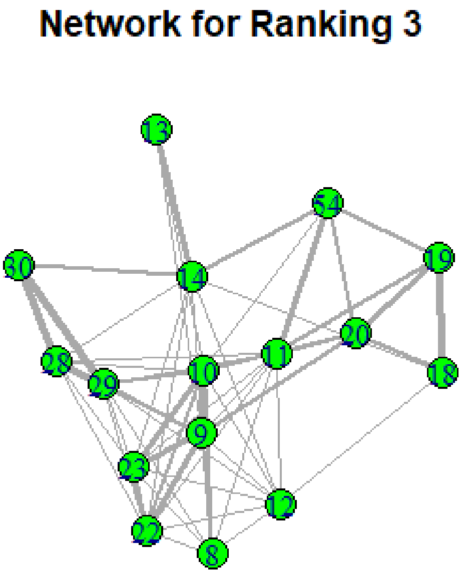
Network for Ranking 1

- Ranking 2 (Cyan)
  Observation: The given network shows a fair amount of connectivity, more than the first ranking.
  Justification: Although the network connectivity is more than ranking 1, I don't agree with the ranking. The gang members in raking 2 are not as involved in co-offending activities as the higher-ranked members.

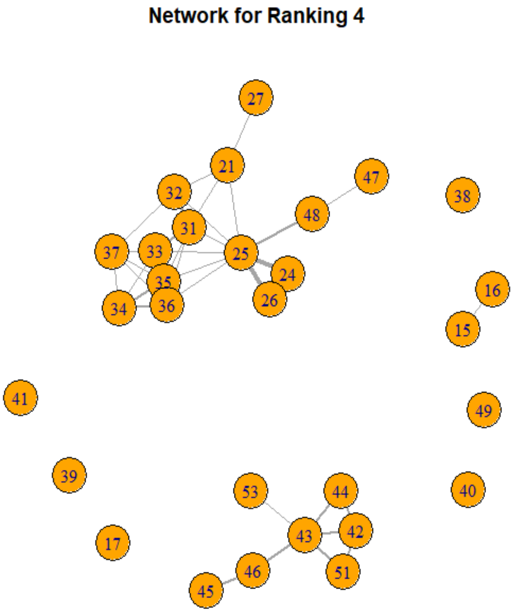


Network for Ranking 2

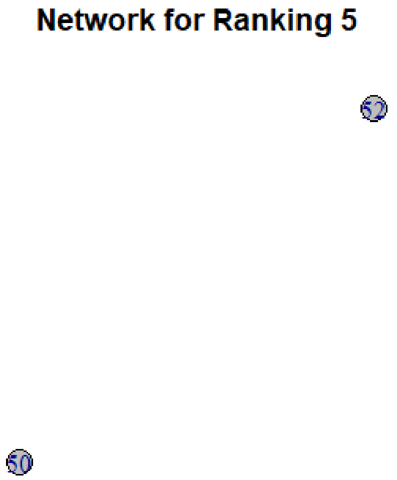
- <u>Ranking 3 (Green)</u>
  Observation: This network shows a higher connectivity than both the previous networks.
  Justification: Gang members in this ranking are likely to be more central in the network which means that they have a higher level of involvement in co-offending activities.


Network for Ranking 3

- <u>Ranking 4 (Orange)</u>
  Observation: This ranking shows more complex structures and connectivity than the other networks.
  Justification: Gang members in ranking 4 are most likely to play a primary role in the network.They also may be heavily involved in serious co-offending activities.


Network for Ranking 4

- Ranking 5 (Grey)
  Observation: As we can see this is a very isolated structure with only 2 grey nodes and no connections at all..
  Justification: In this ranking, the gang members are more likely to be less involved in any co-offending activities.



Overall, I feel like the assigned ranking values do not reflect the seriousness of the co-offending crimes committed by the gang members in each of the networks. Generally, higher-ranked members have more connections but this is not the case here.

Task Six

a.) To create the three networks based on the gang member's ethnicity, first we filter the network to remove links with a weight value equal to 1. Then we can create the subgraphs for interactions between UK gang members and each of the other ethnicities given and finally plot it as shown below: -

- In this network below, we can see that there is a cluster of magenta nodes (UK gang members) and also some isolated blue(West African gang members) and magenta nodes on the periphery. This pattern suggests that there is a primary group with closely interacting gang members who might be in some serious co-offending crimes along with some outliers.

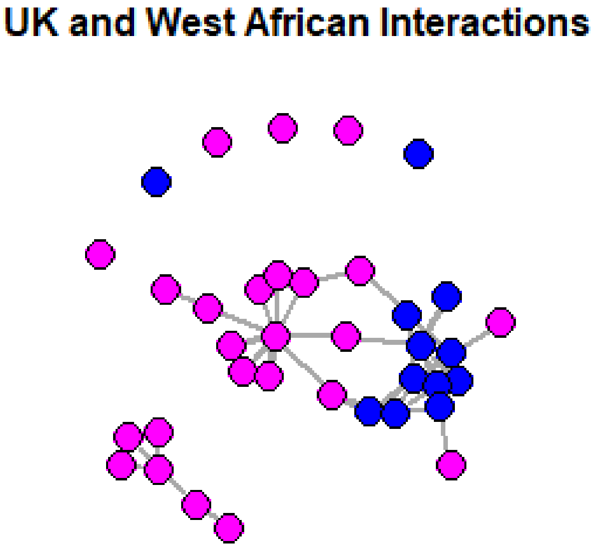


Figure 11: UK and West African Interactions

- In the given network below we can see there is a big cluster of cyan and magenta nodes in the middle with some isolated nodes at the outside of the network. The cyan nodes form a tightly packed cluster as compared to the magenta ones which are more dispersed. This indicates that there is a stronger cohesion among gang members of one group (magenta - UK) with a lot of core and widespread connections whereas the other group (cyan- Caribbean) has fewer gang members (nodes).

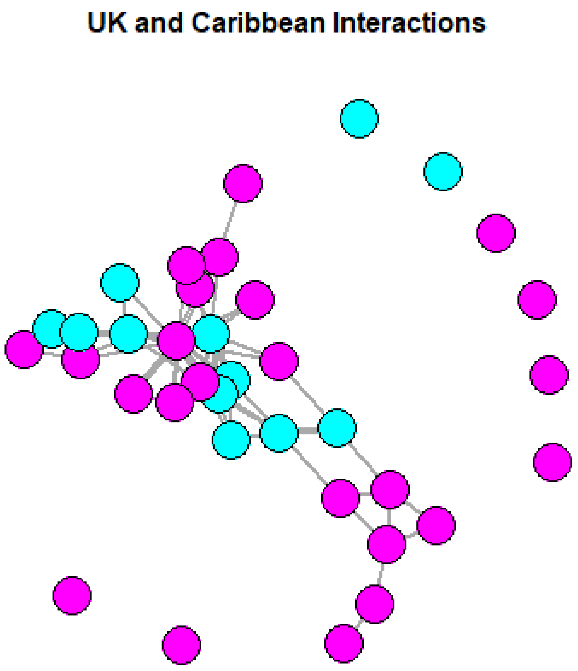


Figure 12: UK and Caribbean Interactions

- As we can see below in the network, there are some magenta nodes forming a loose cluster with some grey nodes scattered around it. There are a lot more gang members of one group (magenta - UK) overall in this network suggesting that they might be the dominating ethnic group with some interactions with the other group(grey- East African) in co-offending crimes. There are also some isolated magenta and grey nodes at the periphery with no connections whatsoever.

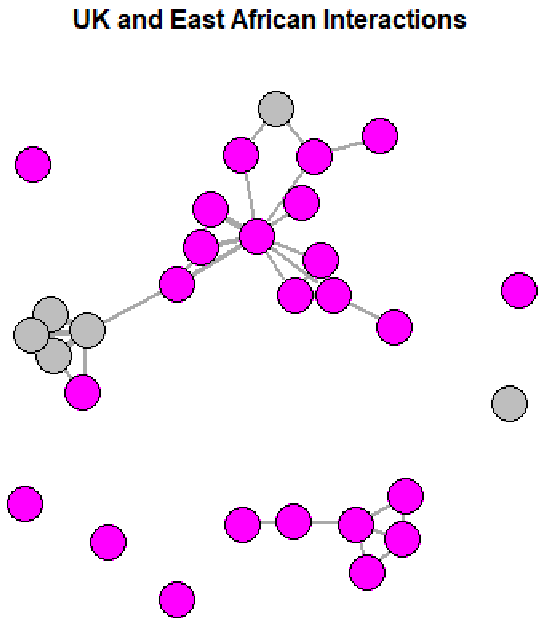
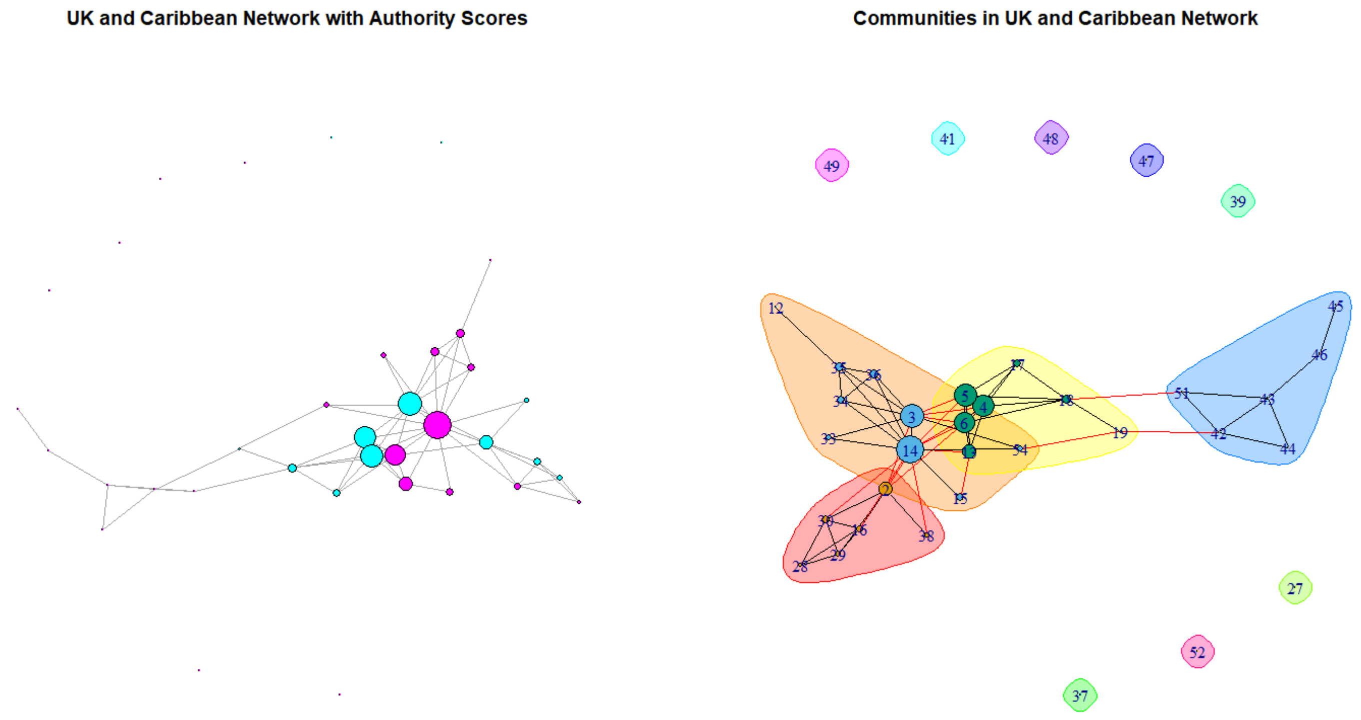


Figure 13: UK and East African Interactions

b.) To calculate the authority scores of the gang members, I used the function *authority_scores* and then set the node size to 10 times the authority score. Then a 2-panel plotting window was created and the first network was plotted. The next step was to identify the communities using the *cluster_optimal()* function to then plot it in the second panel.

The left panel shows the network graph with nodes assigned as authority scores. The authority scores are directly proportional to the node size so the larger the node. So the higher their authority score, the more important they are with a larger influence in the network. This network structure shows a highly interconnected pattern with a lot of edges connecting various nodes, which indicates many co-offending criminal interactions. There are also some very small nodes not connected to the main network which mostly represent the gang members with no co-offending records with the members in the main network.

The network in the right panel visualises the identified communities within the same network. Each of the different colors represent a different community and the numbers represent the size of nodes/ number of members in that particular community. As we can see, there is a presence of different communities which means that there are groups of gang members which have a dense connection and stronger interactions within their respective communities as compared to the rest of the network plot.


UK and Caribbean Network with Authority Scores

Communities in UK and Caribbean Network

Figure 14: Two-panel plot focused on the UK and Caribbean network interactions

c.) Based on all the networks in Task 6 part a and part b, it can be noted that there is evidence to both support and contradict the hypothesis that co-offending occurs mostly among gang members of the same ethnicity which I have explored below: -

1. Evidence supporting the hypothesis:
    - In *figure 11*, we can see that there is a big network of magenta (UK) nodes which are densely clustered together. This indicates that there is some co-offending among gang members of the same ethnicity (UK).
    - Similarly, in *figure 12*, we observe that the purple (Caribbean) nodes have some close clusters indicating that although there are some interactions with the teal nodes, there have been some co-offending crimes of the same ethnicity.
    - In the second plot of *figure 14* there are some distinct communities which seem to be dominated by a single ethnicity (blue and orange communities).
2. Evidence contradicting the hypothesis:
    - The *figure 12* shows a highly- interconnected network between the teal and magenodes indicating that there are inter-ethnic co-offendings.
    - Also in *figure 13*, we can see that the UK and East African communities are inter-connected. These interactions between the purple and grey nodes indicate cross-group interactions.
    - Looking at the second panel plot in *figure 14*, we can note that there are some mixed communities with nodes of varied colors. This contradicts the hypothesis as there are some inner-ethnic interactions.

Conclusion

The analysis of criminal interactions in the UK gang members and their connections with different countries like- the West African, Caribbean, and East African gang members stated complex patterns of co-offending. By identifying the key gang members by centrality measures, the crucial gang members have been highlighted. The influence of certain gang members and distinct communities was emphasised by the calculation of authority scores. The network visualisation and calculation of authority scores provided insights into the structure of the interactions.

We also found that among ethnic communities, Caribbean and West African gang members have the most connections indicating strong inter-ethnic interactions. The evidence supports the hypothesis that co-offending crimes occur mostly among gang members of the same ethnicity which includes the presence of dense clusters of same-ethnicity nodes in some networks. However, there is also evidence of some significant inter-ethnic interactions which indicates that co-offending is not limited to gang members of the same ethnicity.

Overall, we can conclude that ethnicity plays a role in formation of co-offending connections and networks, it is not the only factor. The interactions among gang members are influenced by various factors such as birthplace, number of arrests, and ranking.