# Using R with ArcGIS to find similar groups (clusters) from network connectivity data using social network analysis (SNA) techniques with the *igraph*, *dplyr*, *sp*, and *RColorBrewer* packages

Before proceeding to the example, you must have the following installed on your computer:

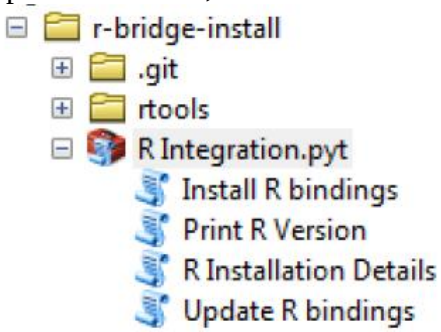
**PREREQUISITES**
ArcGIS 10.3.1 or later

1. R Statistical Computing Software, 3.2.0 or later
   - 32-bit version required for ArcMap, 64-bit version required for ArcGIS Pro (Note: the installer installs both by default).
   - 64-bit version can be used with ArcMap by installing Background Geoprocessing and configuring scripts to run in the background.
2. R ArcGIS Bridge

**SETUP INSTRUCTIONS**
**ArcGIS 10.3.1 or later**

- In the Catalog window, navigate to the folder containing the Python Toolbox, `R Integration.pyt`. *Note*: You may have to first add a folder connection to the location that you extracted the files or downloaded via GitHub.

- Open the toolbox, which should look like this:




Run the `Install R bindings` script. You can then test that the bridge is able to see your R installation by running the `Print R Version` and `R Installation Details` tools.

**Background information**
Network connectivity datasets are a common data type both within and outside the GIS sector. Examples of network datasets are migration of population (e.g. immigration, tourists, business travelers, students, etc.) or import/export of goods from one area to another. These datasets represent flows (edges or links) between two entities (nodes). Social Network Analysis [1] techniques have been developed to use this type of connectivity information and find groups (clusters) of similar entities as well as derive a number of indicators that characterize each entity

and its influence in the network. This tool uses network information (i.e. nodes, links, and directions) of units (e.g. countries, areas, administrative units) and aggregates them into groups (clusters) based on their network connectivity.

The *Network Analysis Grouping* tool uses the *igraph* [2], *dplyr* [3], *sp* [4], and *RColorBrewer* [5] R packages to run FAMD on any combination of quantitative and qualitative variables.

## Explanation of Parameters

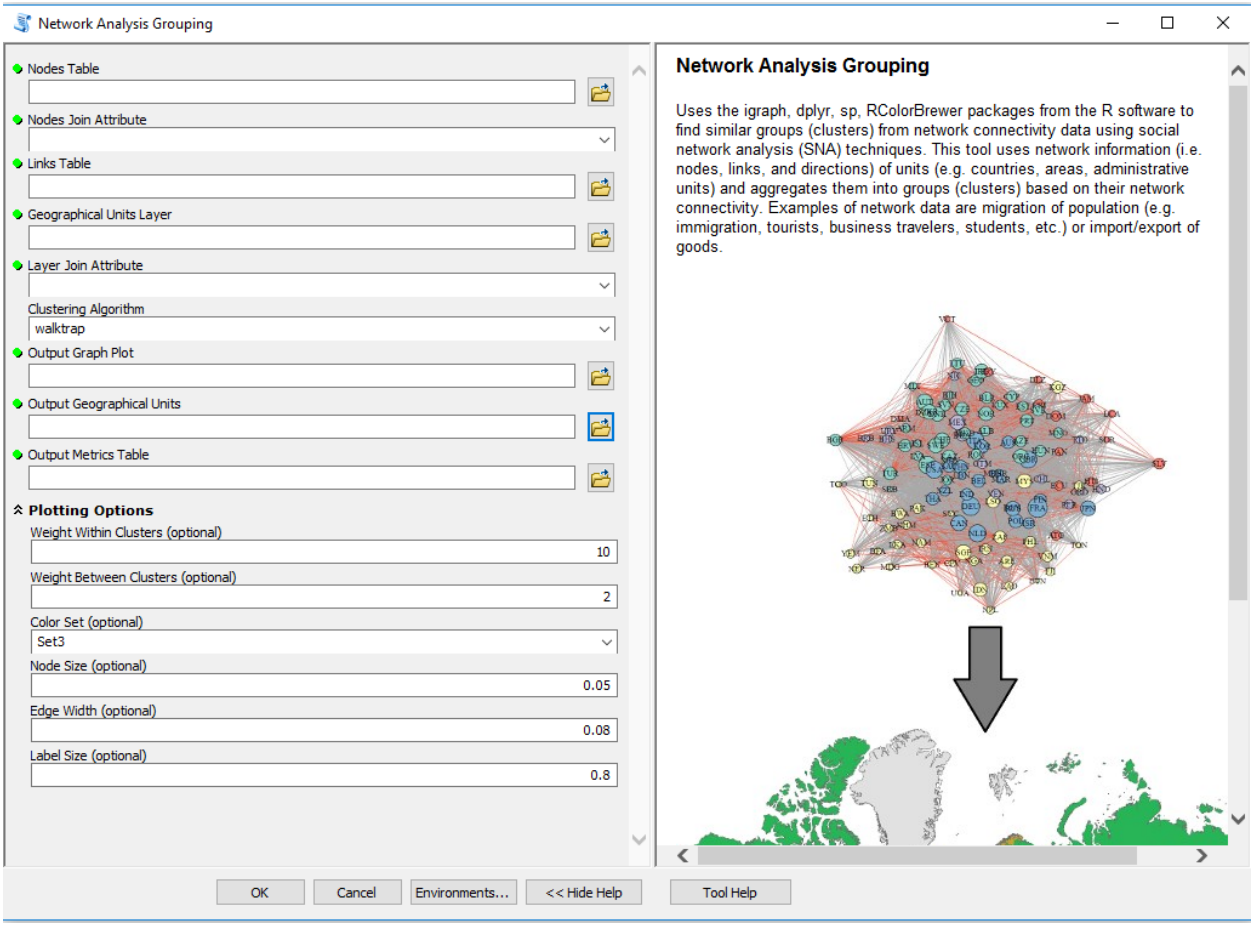



Figure 1. *Network Analysis Grouping* tool.

*Nodes Table*: Input .csv file containing a list of all the units (e.g. countries, administrative units) that are to be considered as "nodes" in the network grouping analysis. This table must include a unique identifier attribute for each record as well as attributes specifying both the outgoing and incoming flows. Flow can be represented by number of people traveling, export/import of goods, energy, etc.

***NOTE: in some case it might be best to use the natural logarithm of total amount of outgoing/incoming flow, especially when flow quantity is very large. Having large numbers will skew the output results so it is recommended to consider natural log as an option.***

*Nodes Join Attribute*: Attribute field from the input nodes table that will be used to join with Geographical Units layer.

*Links Table*: Input .csv file containing a list of all the connections between sending and receiving units (e.g. countries, administrative units) that are to be considered as "links" in the network grouping analysis. This table must include a unique identifier code for both the sending and receiving systems, as well as an attribute specifying the outgoing flow. Flow can be represented by number of people traveling, export/import of goods, energy, etc.

***NOTE 1: make sure the codes used for the sending and receiving units match exactly those used in the nodes input table.***

***NOTE 2: make sure the outgoing flow attribute contains the same values used in the nodes input table. Therefore, if you used natural logarithm in the nodes input table, make sure to use the same here.***

*Geographical Units Layer*: Input feature class representing all the geographical units (e.g. countries, administrative units) represented in the connectivity network.

*Layer Join Attribute*: Attribute field from the Geographical Units Layer that will be used to join with the Nodes Table.

***NOTE: these attribute values will be matched against those found inside the Nodes Table***.

*Clustering Algorithm*: Input algorithm to be used for the grouping analysis.

*walktrap*: detecting community strucure via short random walks. The idea is that short random walks tend to stay in the same community because there are only a few edges that lead outside a given community. Walktrap runs short random walks and uses the results of these random walks to merge separate communities in a bottom-up manner.

*spin_glass*: this function tries to find communities in graphs via a spin-glass model and simulated annealing. In this model, each node can be in one of c spin states, and the edges specify which pairs of nodes would prefer to stay in the same spin state and which ones prefer to have different spin states. Communities can be defined according to the spin states of nodes after running the model.

***NOTE: Refer to R igraph package help page for more information about the algorithms. http://igraph.org/r/doc/communities.html***

*Weight Within Clusters (Optional)*: Number of dimensions selected from the results of the analysis.

*Weight Between Clusters (Optional)*: Input number to ajust the distance between nodes between different clusters.

*Color Set (Optional)*: To select color sets for different clusters. Demos of color sets can be found https://www.nceas.ucsb.edu/~frazier/RSpatialGuides/colorPaletteCheatsheet.pdf

***NOTE: Refer to R RColorBrewer package help page for more information about the color sets. https://cran.r-project.org/web/packages/RColorBrewer/RColorBrewer.pdf***

*Node Size (Optional)*: Input number to adjust size of nodes.

*Edge Width (Optional)*: Input number to adjust the width of edges.

*Label Size (Optional)*: Input number to adjust the size of labels.

*Output Graph Plot*: Output PDF showing a plot in the network space of all the input nodes, their connections (links), colored by the group they belong to, based on the tool results.

*Output Geographical Units*: Output layer consisting of the input geographical units layer plus an added attribute field showing the group (cluster) number each unit belongs to.

*Output Metrics Table*: Output .csv table with network analysis indicators for each input node. The same metrics are also printed in the results message window as well.

**Case study**

The sample data for this tool describes network connections between several countries at the global scale in terms of number of tourists that travel between them. Each record in the nodes table (nodes is a term from network analysis literature) has a unique identifier code tied to it. The table also has two other columns showing the number of outgoing tourists and incoming tourists (***in our sample data we took the natural log of this number to avoid issues with node size in the output results***). The links table shows countries in the tourism network and the amount of tourists heading out from a given country (sender) into a different one (receiver). This table must include a unique identifier code for both the sending and receiving nodes, as well as an attribute specifying the amount of outgoing flow (number of tourists in our case). As you can see, the codes used for the sending and receiving nodes match exactly those used in the nodes input table. The links table shows countries in the tourism network and the amount of tourists heading out from a given country (sender) into a different one (receiver). The outgoing flow attribute contains the same values used in the nodes table, thus if you use natural logarithm in the nodes table, make sure to use the same transformation in the links table. We also provide a global shapefile representing countries that will be used in the analysis (Fig. 2).

Figure 2. Countries shapefile that are used as the input geographical layer units in our case study example.

The goal of this case study example is to run network analysis grouping techniques to find similar clusters of nodes (countries) that are most similarly connected to each other. In other words, which countries exchange similar amount of tourists (hence are most connected) that travel between them.

As an input nodes table, browse to ./data/Tourism/nodes.csv file provided inside the CHANS tools repository on Github. As a join attribute, type in 'CODE' (no quotes) which corresponds to the attribute inside the nodes.csv file that contain a unique identifier for each country. For the links table input, browse to ./data/Tourism/links.csv. The geographical units layer in this example will be the global country shapefile provided inside the same data folder. The layer join attribute needs to be a unique identifier in the shapefile attribute table that corresponds to those found inside the nodes.csv table. It does not matter if you have more or less units (countries in our case) in the shapefile, as long as the identifiers match the data type of those found in the nodes table. This attribute is very important as it will be used to join the results of the network analysis clusters into the shapefile to be able to spatially visualize the distribution of the similar groups (by color). In our example, type 'ISO_3_CODE' (no quotes) into the layer join attribute field. You can leave the default settings for clustering algorithm and any plotting graphical parameters provided. Feel free to read the help page of the tool and explore the effect of each one of them at your own convenience. The final step is specifying the three output files needed for this tool to run. The graph plot is a PDF containing the network grouping plot with the results from the network analysis. This plot gives you a better idea of which units have stronger connections (links) with other units in the network and the size of the node circles is proportional to the importance and centrality within it. Finally, specify the output location and name of the result feature class coming out of the network analysis and an output location/name for the .csv file that will contain some of the network analysis metrics for each unit (country). The same metrics are printed out in the geoprocessing message window but it will help having it stored

inside a physical table as well for further inspection. Figure 3 shows an overview of the tool GUI filled out with the parameters as discussed above.

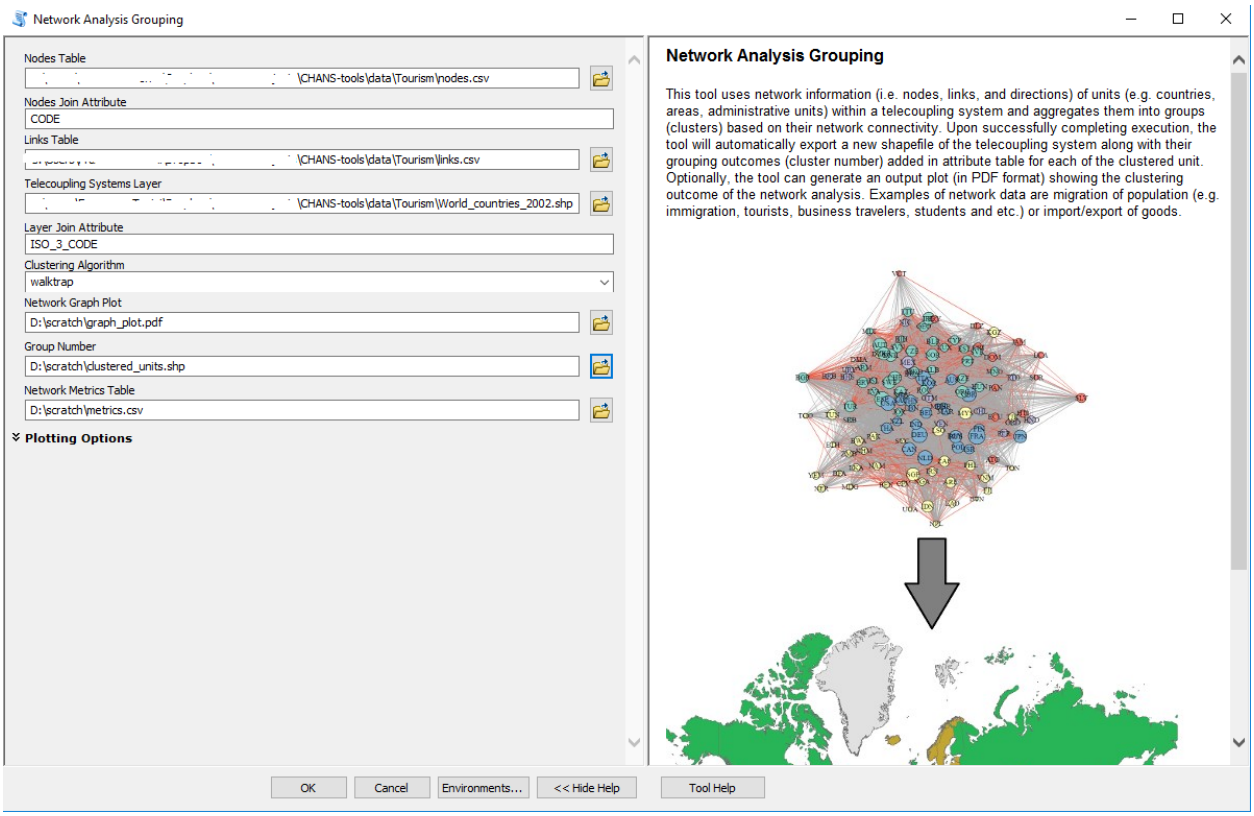



Figure 3. *Network Analysis Grouping* tool along with parameter values filled out for the case study in this tutorial.

If all goes smoothly, the tool identifies groups of most similar units in terms of their network connectivity. If you open the result window to inspect the results (Menu > Geoprocessing > Results), under “Current Session”, right-click the “Message” icon and select “View”. A separate window should open showing you the entire process executed by the tool and all the network metrics for each input node (country), as shown in Fig. 4.

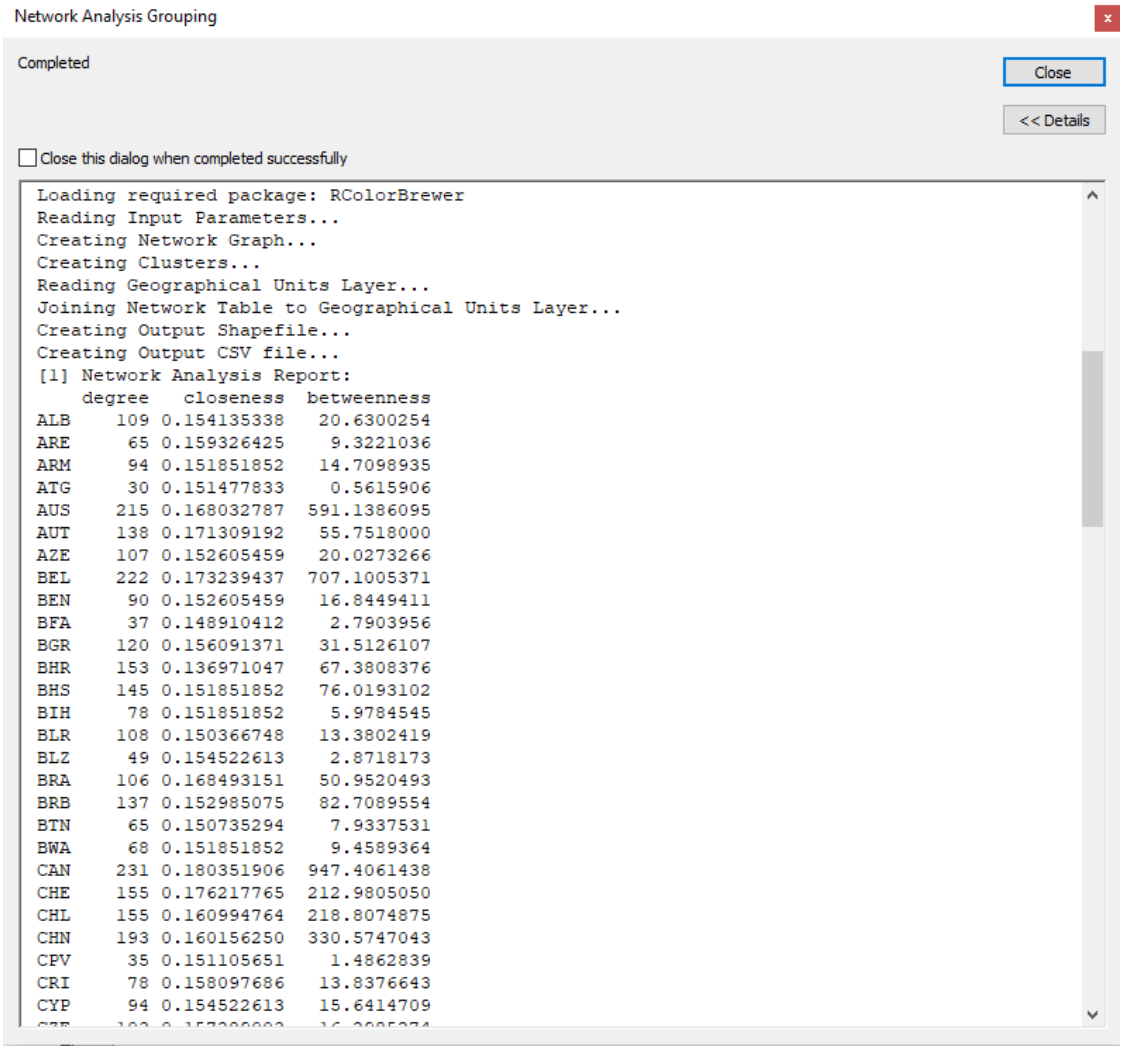



Figure 4. Snapshot of the Messages as a result of the geoprocessing tool execution.

Inside ArcMap, you should now also see the output shapefile colored by group number (Fig. 5). As you can see, countries colored the same implies a strong connectivity in terms of tourists that traveled between them, thus are similarly connected in the network space. The shapefile attribute table shows the new attribute (*cluster_N*) that was joined to it and represent the group number each node/unit belongs to.

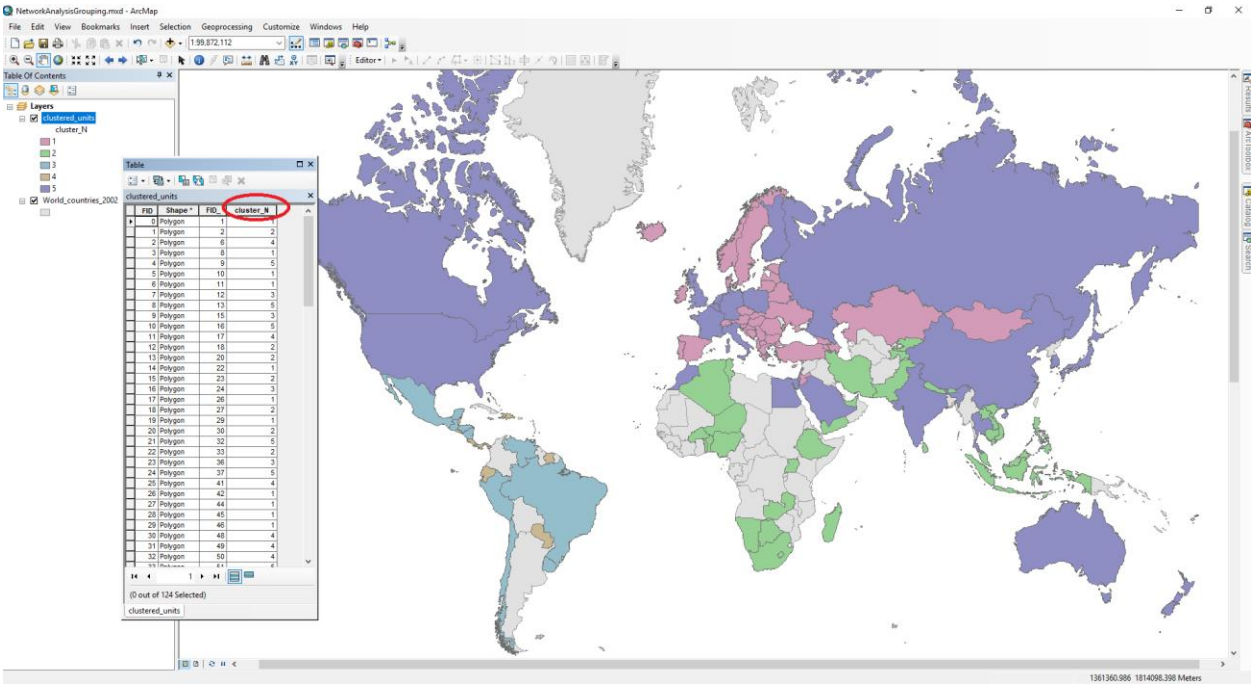



Figure 5. Network groups (same color) of countries with strong connectivity in terms of number of tourists that traveled between them.

If you open the output graph plot PDF, you should see nodes and links represented in the network (not geographical) space (Fig. 6). Colors identify similar groups of units and node size is directly proportional to the importance within the network. These colors are not meant to be the same as those shown in the output map and you can control all graphic parameters from the optional GUI parameters to customize your needs. However, you can notice a one-to-one correspondence between countries that were shown to belong to the same group in the map and here in the network space.

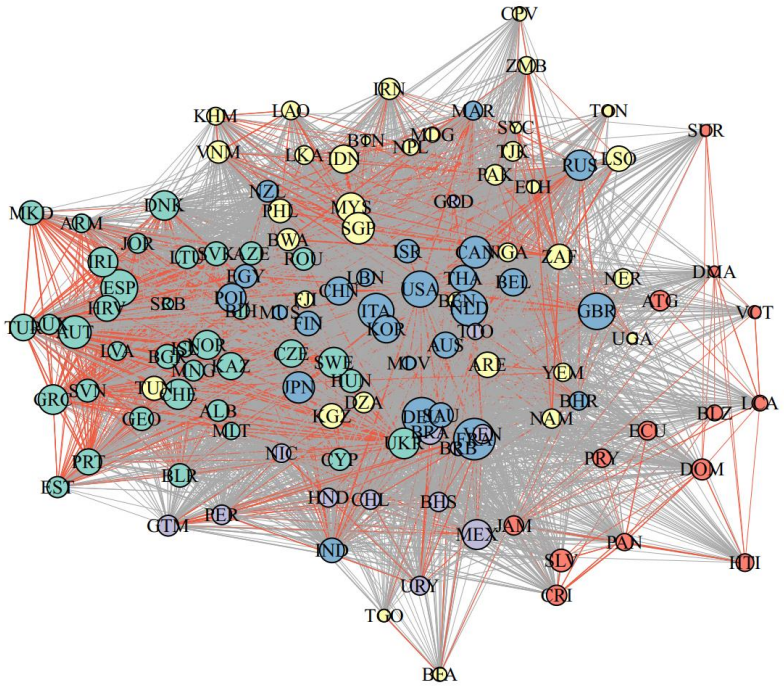



Figure 6. Plot of network groups (same color) of countries with strong connectivity in terms of number of tourists that traveled between them. This output graph is saved into the output PDF file chosen in the tool GUI.

Finally, the same network metrics that are printed out in the geoprocessing message window (Fig. 4), are saved into the output .csv file for you to access them and make further conclusions about each node in the network. Currently, the metrics chosen for the output are *degree*, *closeness*, *betweenness* and are three of the most commonly used network metrics found in the literature.

*Degree*: Importance score based purely on the number of edges (links) held by each node. A higher number means a more connected node. It is often considered a measure of direct influence.

*Closeness*: Scores each node based on their 'closeness' to all other nodes within the network. It calculates the shortest paths between all nodes, then assigns each node a score based on its sum of shortest paths. It tells how many direct, 'one hop' connections each node has to other nodes within the network. Useful for finding very connected nodes who are likely to hold most information or nodes who can quickly connect with the wider network. Higher values mean less centrality.

***NOTE: Closeness centrality can help find good 'broadcasters', but in a highly connected network you will often find all nodes have a similar score.***

*Betweenness*: measures the number of times a node lies on the shortest path between other nodes. This measure shows which nodes act as 'bridges' between nodes in a network. It does this by identifying all the shortest paths and then counting how many times each node falls on one. Useful for finding the nodes that influence the flow around a system. A high betweenness count could indicate someone holds authority over, or controls collaboration between, disparate clusters in a network.

| | degree | closeness | betweenness |
|---|---|---|---|
| ALB | 109 | 0.154135338 | 20.63002545 |
| ARE | 65 | 0.159326425 | 9.322103636 |
| ARM | 94 | 0.151851852 | 14.7098935 |
| ATG | 30 | 0.151477833 | 0.56159059 |
| AUS | 215 | 0.168032787 | 591.1386095 |
| AUT | 138 | 0.171309192 | 55.75180001 |
| AZE | 107 | 0.152605459 | 20.0273266 |
| BEL | 222 | 0.173239437 | 707.1005371 |
| BEN | 90 | 0.152605459 | 16.84494106 |
| BFA | 37 | 0.148910412 | 2.790395562 |
| BGR | 120 | 0.156091371 | 31.51261066 |
| BHR | 153 | 0.136971047 | 67.38083755 |
| BHS | 145 | 0.151851852 | 76.0193102 |
| BIH | 78 | 0.151851852 | 5.978454459 |
| BLR | 108 | 0.150366748 | 13.38024193 |
| BLZ | 49 | 0.154522613 | 2.871817264 |
| BRA | 106 | 0.168493151 | 50.95204933 |
| BRB | 137 | 0.152985075 | 82.70895544 |

Figure 7. Snapshot from the output .csv table showing network metrics for each node and their values.

**References and further reading**


[1] Scott J, Social Network Analysis: A Handbook, 2nd ed. (Sage Publications, London, 2000)

[2] Csardi G, Nepusz T: The igraph software package for complex network research, InterJournal, Complex Systems 1695. 2006. http://igraph.org

[3] Wickham H, Francois R, Henry L, and Müller, K. 2017. dplyr: A Grammar of Data Manipulation. https://CRAN.R-project.org/package=dplyr

[4] Bivand R, Pebesma E, Gomez-Rubio V. 2013. Applied spatial data analysis with R, Second edition. Springer, NY. http://www.asdar-book.org/

[5] Neuwirth E. 2014. RColorBrewer: ColorBrewer Palettes. R package version 1.1-2. https://CRAN.R-project.org/package=RColorBrewer