*Week 8 Group Project - Network Visualization*

Siddhartha Seksaria, Kasey Kane, and Seth Howells

Data Visualization, Fall 2021

Lewis University

Romeoville, IL, USA

[siddharthaseksaria@lewisu.edu](mailto:siddharthaseksaria@lewisu.edu), [kaseytkane@lewis.edu](mailto:kaseytkane@lewis.edu), seththowells@lewisu.edu

# I. Introduction

The graph analytics market was valued at $600 million in 2019, and it is forecasted to grow to $2.4 billion by 2024. This growth is forecasted by the ability to incorporate many different data sources and mine new relations within datasets at ease [1]. A graph, or network diagram, shows the interconnectedness between a set of entities. It does so by employing nodes, or vertices, where connections are represented by edges. Studying graphs have many applications within spheres such as what is studied in this report: Delta airline analysis, social network analysis, and network classifications for wines.

Gephi and Orange3 are open source software tools that can be used to visualize graphs and study connections within a network [2,3]. Each software is able to employ layout algorithms of a network to make extraction of information about the network easier. Examples of the layout algorithms include GeoLayout, Force Atlas 2, Fruchterman Reingold and Circle Pack [4,5,6,7]. After choosing a layout, statistics such as modularity, network diameter, and clustering coefficient can be generated for a network to provide information about the network.

The future sections of this report give a description of the data used, a network description, and network visualization of three different graph datasets. Section II focuses on the flight patterns within Delta hubs, Section III focuses on the analysis of the target word ‘coronavirus’ within Gephi’s Twitter API, and Section IV focuses on the network classification of three cultivars of wine from Italy.

# II. Geo Graph Visualization for Airlines

1. Data Description

An airline representation was collected from the Delta Airlines website, where an edge table was created from each of the US hubs to its destination airports [8]. The US Delta hubs include Atlanta, Detroit, Los Angeles, Minneapolis, New York’s JFK and LGA, Salt Lake City, and Seattle. Every node is representative of an airport and each edge is representative of a flight between two airports. Each edge is undirected, with the assumption that a flight can run from either city. An example of an edge within the edge table is shown below in Table 1.

Table 1. Delta Airline Data - Edge Table Example

| **Souce** | **Target** | **Type** | **Id** | **Weight** | **State** | **Country** |
| --- | --- | --- | --- | --- | --- | --- |
| Atlanta | Orlando | Undirected | 1 | 1.0 | FL | USA |

Gephi offers a GeoLayout plugin to display graphs based on geocoded attributes. In this case, to generate the layout, after the edge table was imported, the auto-generated node table created within Gephi was exported. The latitude and longitude of each airport was added from the Global Airport Database into the node table [9]. This new node table was imported back into Gephi, and GeoLayout employed the use of the coordinates of each node. An example of the imported node table including coordinates is shown below in Table 2.

Table 2. Delta Airline Data - Nodes Table Example

| **Id** | **Latitude** | **Longitude** |
| --- | --- | --- |
| Atlanta | 33.6407 | -84.4277 |

The statistics of the network generated is displayed below in Table 3.

Table 3. Network Statistics of Airline Data

| **Metric** | **Value** | **Description** |
| --- | --- | --- |
| Nodes | 226 | Total number of nodes in the dataset |
| Edges | 580 | Total number of edges in the dataset |
| Average Degree | 5.133 | Average number of edges connected to a node |
| Network Diameter | 3 | The shortest distance between the two most distant nodes |
| Modularity | 0.259 | Sum over all clusters of the number of edges in the cluster minus the number of edges expected by chance in the cluster |
| Avg. Clustering Coefficient | 0.933 | How well connected the neighborhood of the node is |
| Avg. Path Length | 2.201 | Average shortest path between two nodes |

1. Network Description and Visualization

While GeoLayout provides an algorithm to arrange the nodes according to their geographical coordinates, there are also various ways to project the data with GeoLayout. This includes Mercator, transverse Mercator, sinusoidal, and Miller cylindrical. The most well known map projection in cartography is Mercator, and as a result, that is the one utilized in the following visualizations. When utilizing GeoLayout, the edges present are not representative of flight patterns taken. For example, when flying from the USA to Japan, an airline might choose to fly west, over the Pacific Ocean. However, in this representation, a straight line is drawn in two dimensional space from each airport.

Betweenness centrality, a measure of the number of paths that go through a node, provides a representation of the importance of certain airports within a network. This is pertinent to airline data because it can show weaknesses within an airline network. Being overly reliant on one hub or airport can mean disaster if problems arise within that airport. As shown in Fig. 1 below, the Atlanta airport (ATL) provides the highest betweenness centrality of any Delta hub. It has several edges that are weighted of high importance. Minneapolis (MSP) and New York’s JFK airport provide higher centrality than the surrounding airports, however, this is much less than ATL. Within this network, the extreme reliance on ATL as an airport demonstrates fragility in Delta’s operations.

\*blurry from Word – I definitely think I want to turn this in after I export it to Word so the images aren’t blurry. It bothers me



Figure 1. Network nodes and edges colored by betweenness centrality. The highest betweenness centrality comes from Atlanta (ATL), Minneapolis (MSP), and New York (JFK). The closer to black the color of the node, the higher the betweenness centrality.

Modularity, a measure of the strength of division in a graph, can be examined within airline networks. The interconnectedness of flight networks can be broken up into modules, where there are more flights within a module than flights outside of that module. Within this network, as shown in Fig. 2, four modules appear, two of which, colored in pink and light green, appear at a higher percentage than the rest. The light green module is centered around the Atlanta hub, a node with 179 degree connections, the highest within the dataset. Minneapolis and Salt Lake City hubs contribute to the connections within the pink module, New York’s JFK hub contributes mainly to the blue module, and Seattle and Los Angeles contribute to the dark green module. The modularity speaks to the way in which airlines function, where this is representative of a ‘hub and spoke.’ Each of Delta’s hubs have many flights to subsidiary or smaller airports that are within its module. If Delta operated within a ‘point-to-point’ method of air travel, the modules would be hard to differentiate because there would not be a strength in division of certain airports choosing to fly from a spoke to a hub or from hub to hub to spoke [10]. Delta’s operations are heavily reliant on hubs, with an emphasis on Atlanta, and this is visualized with Fig. 2 focusing on the modularity of the network.

\*Label hubs in figure — makes it blurry bc of pasting this from Word

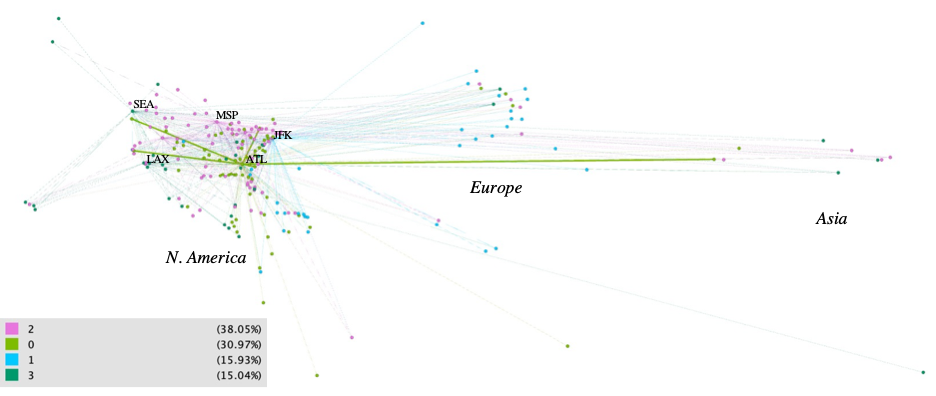


Figure 2. Network nodes and edges colored by modularity. Gephi calculated the modularity to be 0.259 and broke the network into four modules, colored pink, light green, blue, and dark green.

# III. Twitter Tweets and Hashtags using Graph

1. Data Description

Data contained in this section was generated through Gephi’s Twitter API. The API was focused on collecting Tweets that contained the target word: Coronavirus. Tweets that contained the target word as well as a hashtag were obtained. Hashtags that successfully passed the filter became the nodes of the network analysis. The edges pertain to the relationship of the original Tweet and the reply. For example, a Tweet from the CDC with a hashtag #coronavirus can be responded to with another hashtag like #getvaccinated. The total number of responses with the same hashtags (i.e. #coronavirus → #getvaccinated) is represented as the weight.

The purpose of this analysis is to examine the relationship of social data concepts linked to coronavirus in efforts to identify social networking communities and attributes of that community. Social data concepts refer to the interactions in metadata tags, i.e. #hashtag, to identify common themes in the community [11]. Furthermore, edge weights were applied to visualize the relationships between the largest node in the community with other communities. Examples of the node and edge table are found below in Tables 4 and 5, respectively.

Table 4. Twitter Data - Nodes

| **Id** | **Label** | **twitter\_type** |
| --- | --- | --- |
| #covid19 | #covid19 | Hashtag |

Table 5. Twitter Data - Edges

| **Souce** | **Target** | **Type** | **Id** | **Weight** |
| --- | --- | --- | --- | --- |
| #pandemic | #maskup | Undirected | 3 | 4.0 |

1. Network Analysis

Several tests were performed to analyze the network, which are found in Table 6. Modularity class is responsible for determining the number of communities and assigning each node to the corresponding community. To illustrate the communities, the node color was set to the appropriate community label. The node size reflects the weighted degree so that hashtags with the largest amount of interactions are the largest nodes. The edges were colored by the source node color while the thickness of the curved lines represents the weighted edge so that high interactions between the nodes are thicker.

The network visualization algorithm utilized was Circle Pack. This algorithm packs similar nodes based on hierarchical structure, where the first hierarchy dictates the largest influence. Circle Pack layout uses Mike Bostock’s circle packing algorithm.

A hierarchy was used to group the graph into communities. The hierarchy had the following structure:

* Hierarchy 1: Modularity Class
* Hierarchy 2: Clustering Coefficient
* Hierarchy 3: PageRank
  + Epsilon = 0.001
  + Probability = 0.85

Partition of Nodes

* Identify node communities by color through their respective Modularity Class

Ranking Nodes

* Larger nodes are ranked higher and smaller nodes are ranked lower

Ranking Edges

* Thicker edges indicate a high Degree, while thin edges indicate low Degree

Table 6. Network Statistics of Twitter Data

| **Metric** | **Value** | **Description** |
| --- | --- | --- |
| Average Degree | 11.987 | Average number of edges connected to a node |
| Avg Weighted Degree | 38.0734 | Average sum of weights of the edges of nodes |
| Modularity | 0.733 | Sum over all clusters of the number of edges in the cluster minus the number of edges expected by chance in the cluster |
| Avg. Clustering Coefficient | 0.831 | How well connected the neighborhood of the node is |
| Avg, Path Length | 2.891 | Average shortest path between two nodes |

1. Network Visualization

The Twitter hashtag network produced 149 nodes (hashtags) and 893 edges (relationships), which resulted in nine unique communities (color). The closer the communities are to the target word, #Coronavirus, the closer the relationship as determined by PageRank. Of the nine communities, five were found to possess anti-vaccine sentiment, engage in conspiratorial concepts, and contain politically charged hashtags. The remaining four communities (blues, purples) show signs of collective concepts, pro-vaccination stances, and attentiveness towards COVID variants.

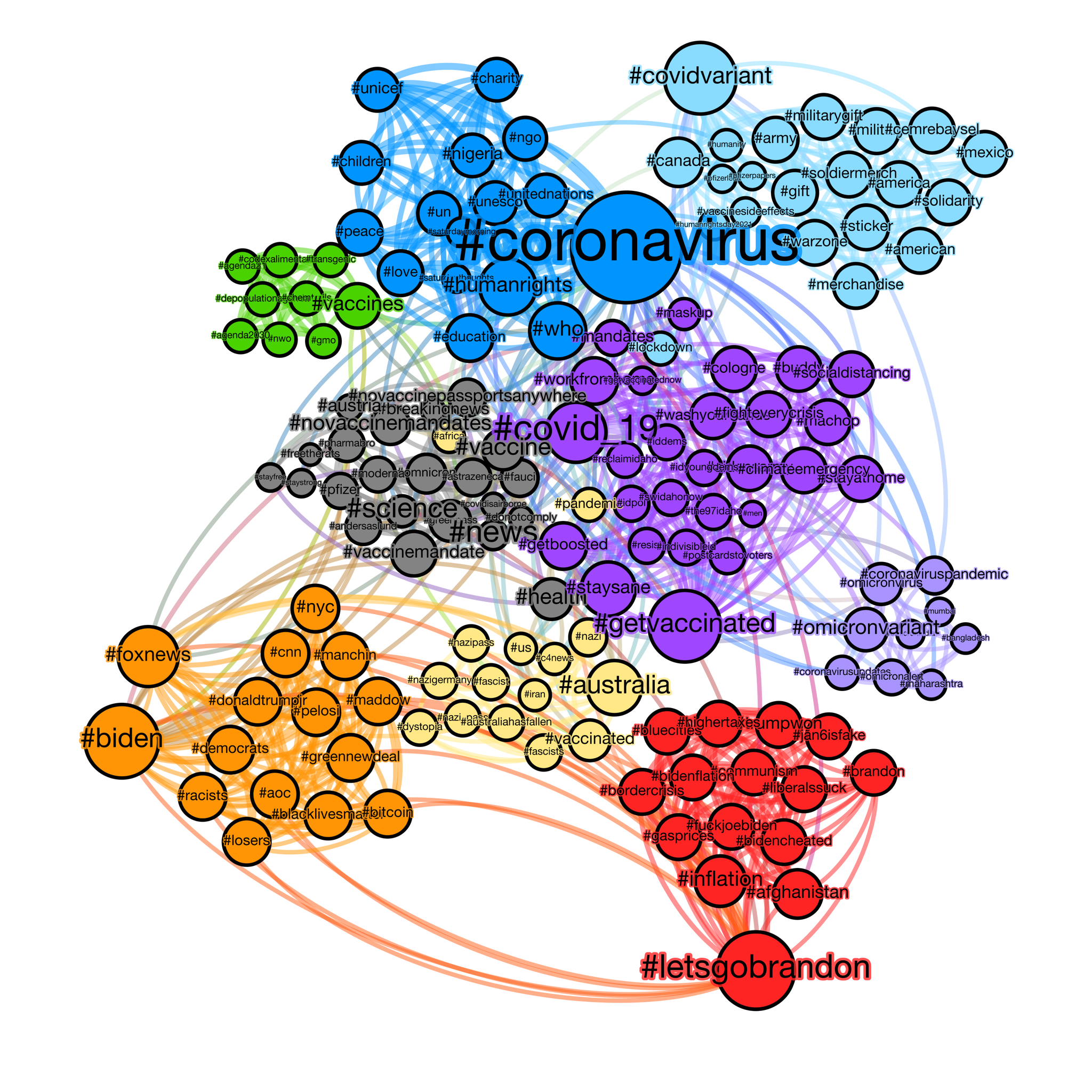


Figure 3. Twitter COVID Hashtag network. Circle Pack layout algorithm applied to social data network to visualize the Communities (node color). Layout algorithm ranked the Hierarchy as follows: Modularity Class (H1), Clustering Coefficient (H2), and PageRank (H3). Edge color and thickness is represented by source node color to target node.

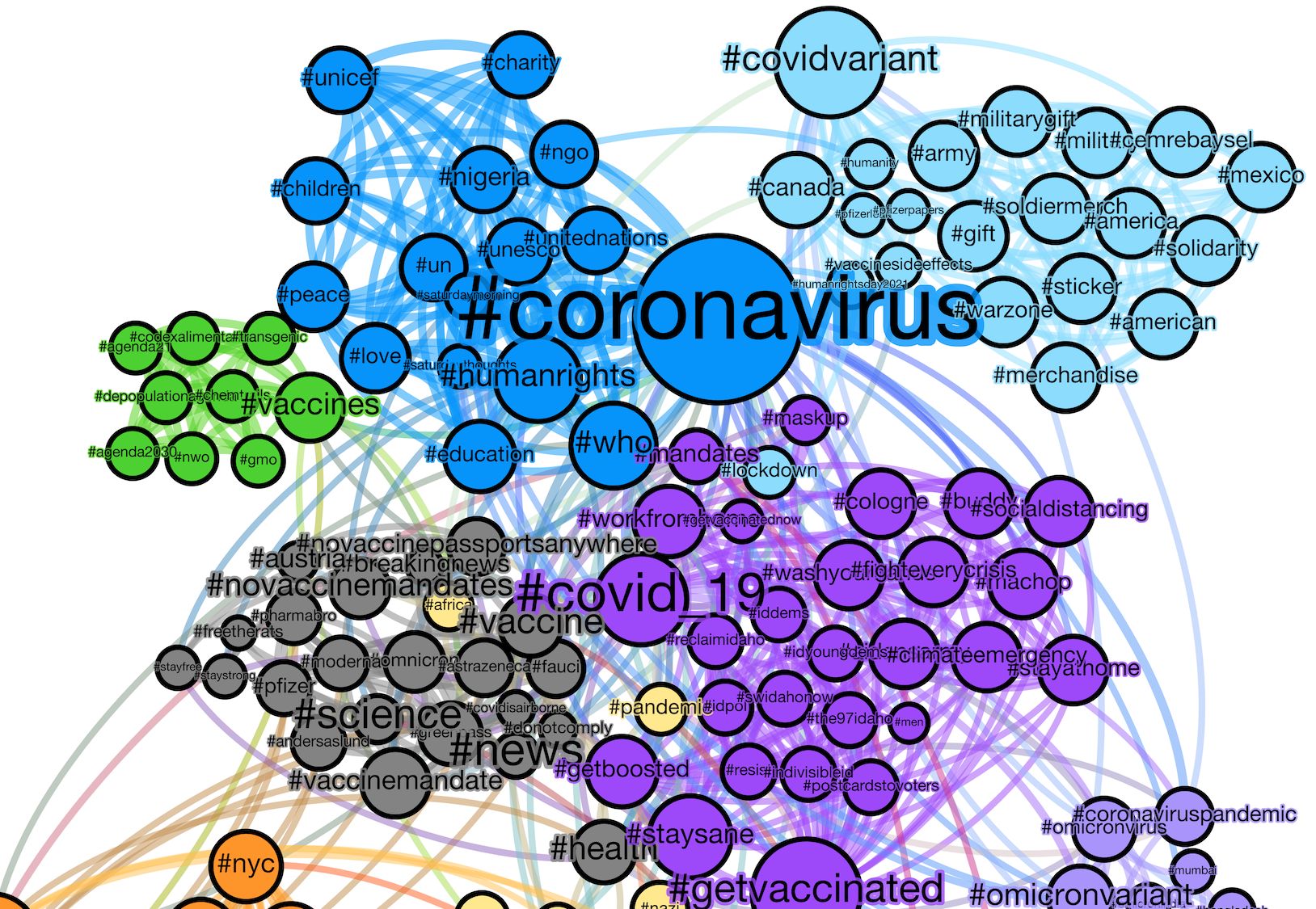
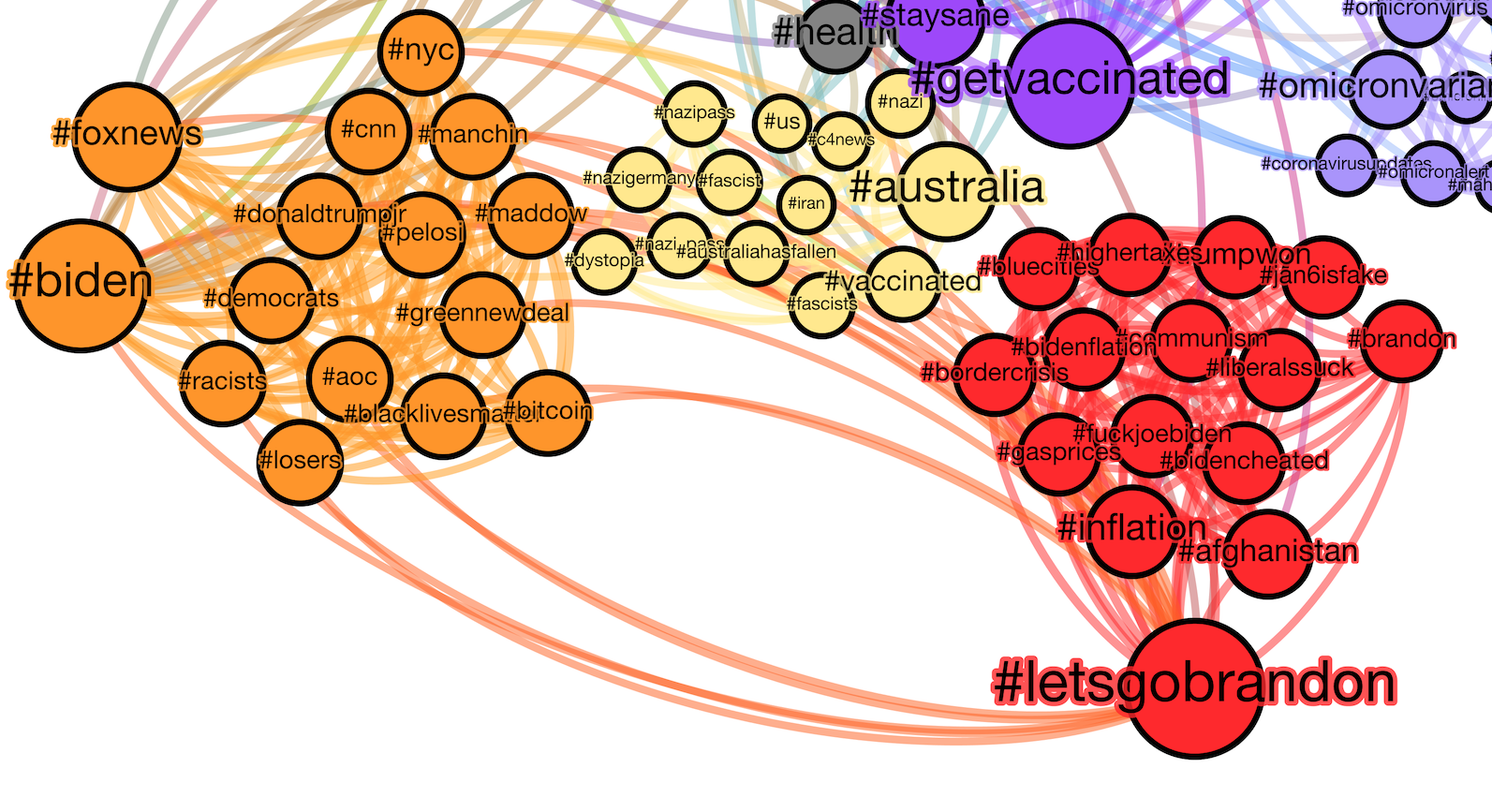


Figure 4. Top Half: Green and Grey communities show signs of COVID-19 vaccine non-compliance and interactions with conspiratorial concepts (i.e. #NWO, #donotcomply, #depopulationagenda). The other (3) contradistinctive communities show signs of pro COVID-19 vaccination policies and display a high degree of collective concepts (i.e. #UN, #WHO, #humanrights).

Figure 5. Lower Half: Orange and Red communities heavily engage in Tweets with politically charged hashtags. The edge structure in both of these communities display a high degree of intercommunity interactions, yet a low degree of outbound edges toward a diverse set of communities, leaving room to speculate “echo chambers.”

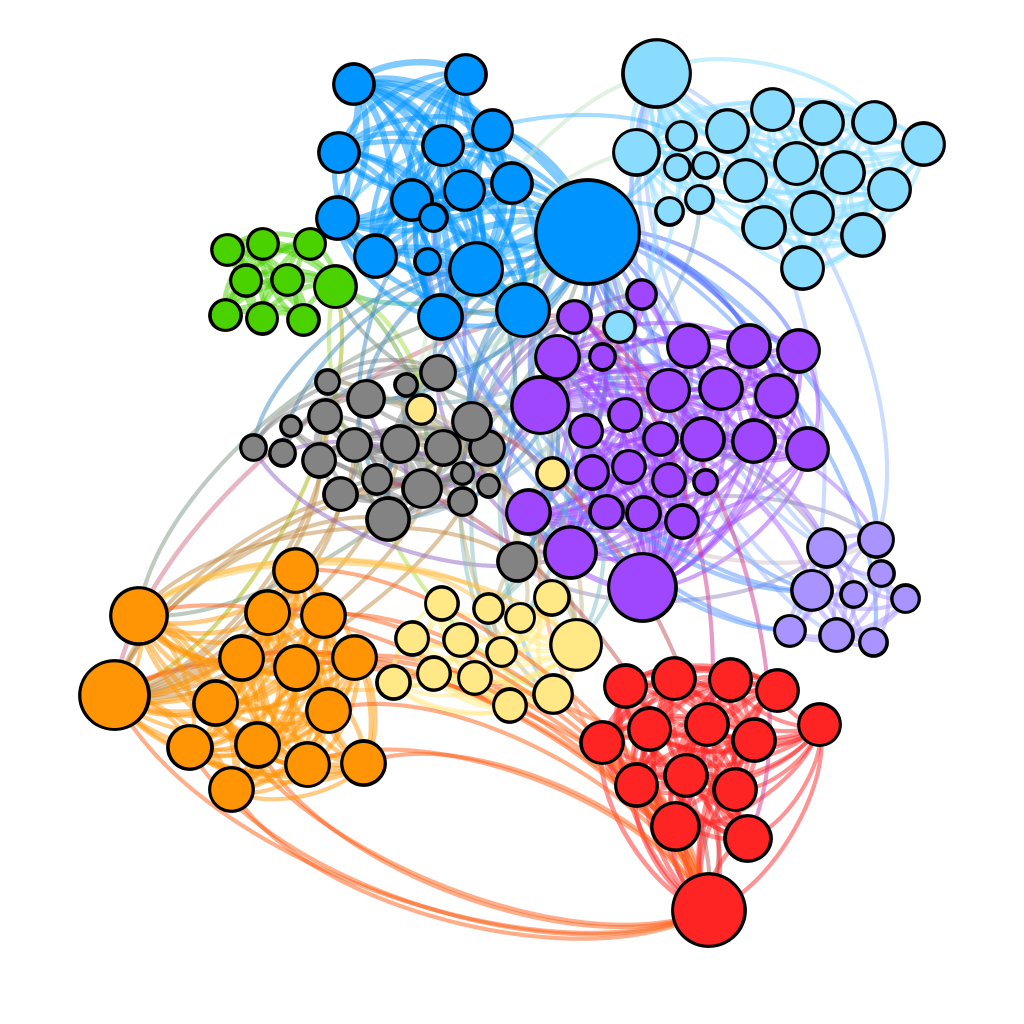


Figure 6. Total Communities with no labels.

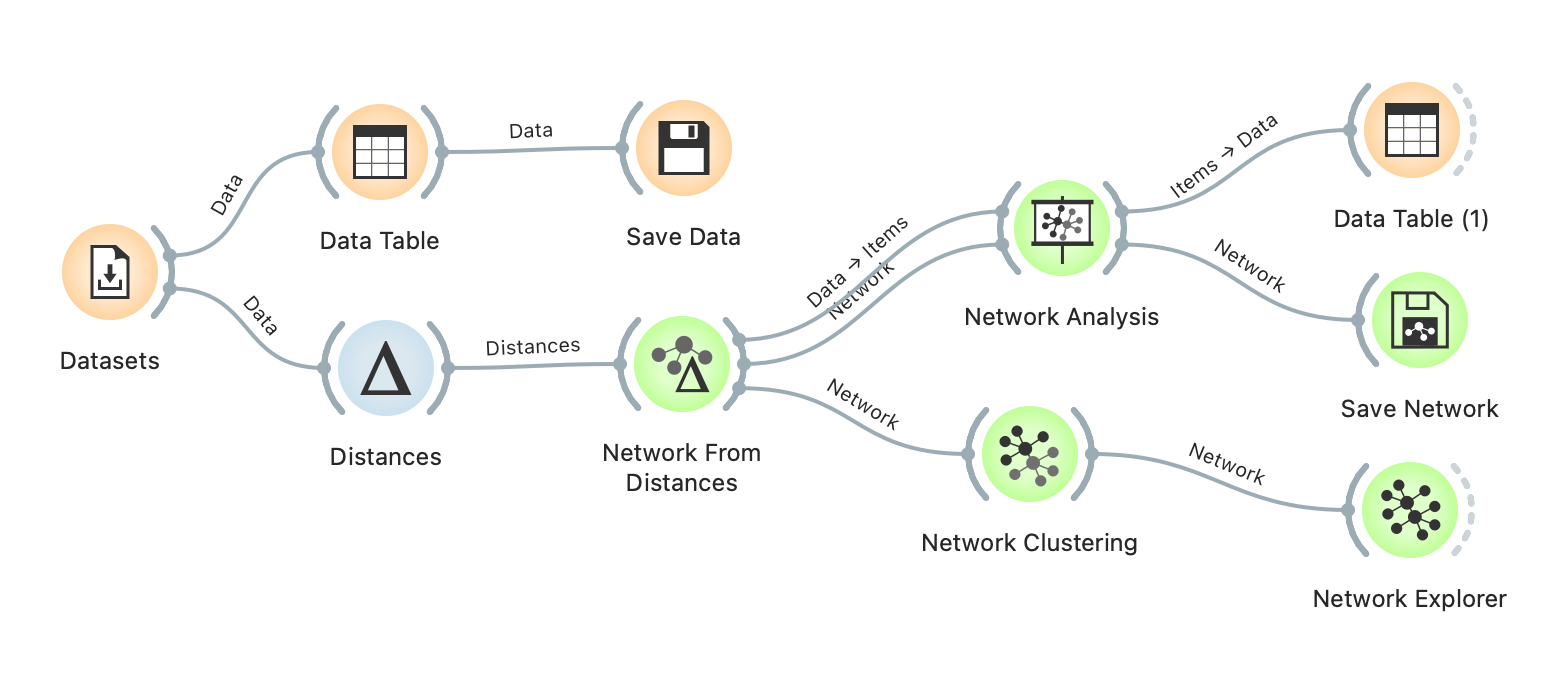
# IV. Network Classifications using Orange3

1. Data Description

Wine data set from UCI [12]. The data are the results of a chemical analysis of wines grown in the same region in Italy but derived from three different cultivars. The analysis determined the quantities of 13 constituents found in each of the three types of wines. The 13 different data attributes are Alcohol, Malic Acid, Ash, Alkalinity of Ash, Magnesium, Total phenols, Flavonoids, Non Flavonoid phenols, Proanthocyanidins, Color intensity, Hue, OD280/OD315 of diluted wines and Proline. The dataset is well balanced and should be able to use the distance of each row in a network classifications model.

1. Network Analysis

Wine Dataset is available using Orange3. Using the Distance widget, the Euclidean distance is calculated between each row. The network from Distance widget helps generate a graph dataset with a node as each sample and edges between the nodes is the Euclidean distance of each one of them. The smaller the distance value, the similar the wine or dataset is and likely belong to the same cultivars. Larger the distance, the Wine will likely be from different cultivars.

Figure 7. Orange dataflow used to generate network data.

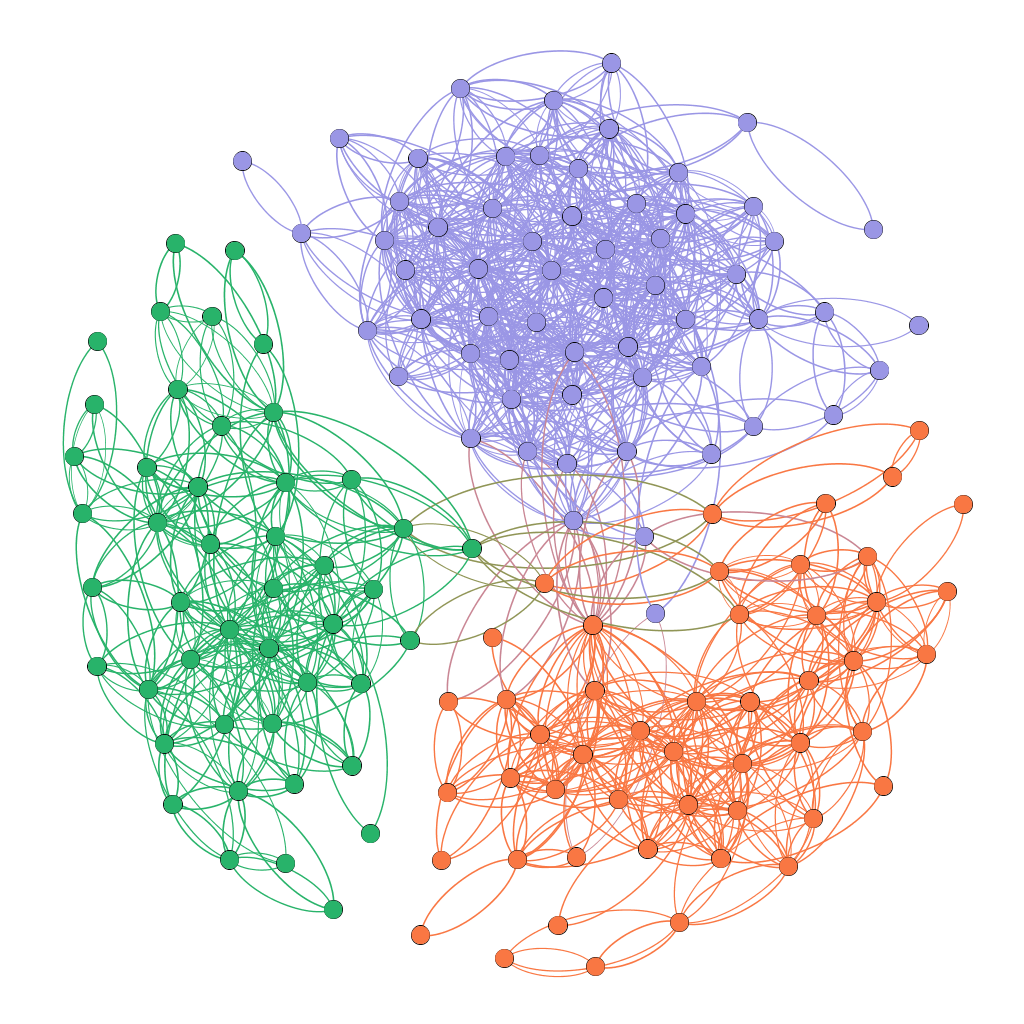
The statistics of the network generated is displayed below in Table 7.

Table 7. Network Statistics of Wine Data

| **Metric** | **Value** | **Description** |
| --- | --- | --- |
| Nodes | 155 | Total number of nodes in the dataset |
| Edges | 1727 | Total number of edges in the dataset |
| Average Degree | 22.28 | Average number of edges connected to a node |
| Network Diameter | 12 | The shortest distance between the two most distant nodes |
| Modularity | 0.584 | Sum over all clusters of the number of edges in the cluster minus the number of edges expected by chance in the cluster |
| Avg. Clustering Coefficient | 0.524 | How well connected the neighborhood of the node is |
| Avg. Path Length | 6.969 | Average shortest path between two nodes |

1. Network Visualization

The Fruchterman Reingold force algorithm is a good way to show closely related nodes. The modularity algorithm resulted in three different communities, which reflected the number of cultivars the data was collected. The node color represents each of the cultivars as shown below in Fig. 8. This helps in visualizing all the nodes in the graph.

Figure 8. Fruchterman Reingold layout of wine dataset. 

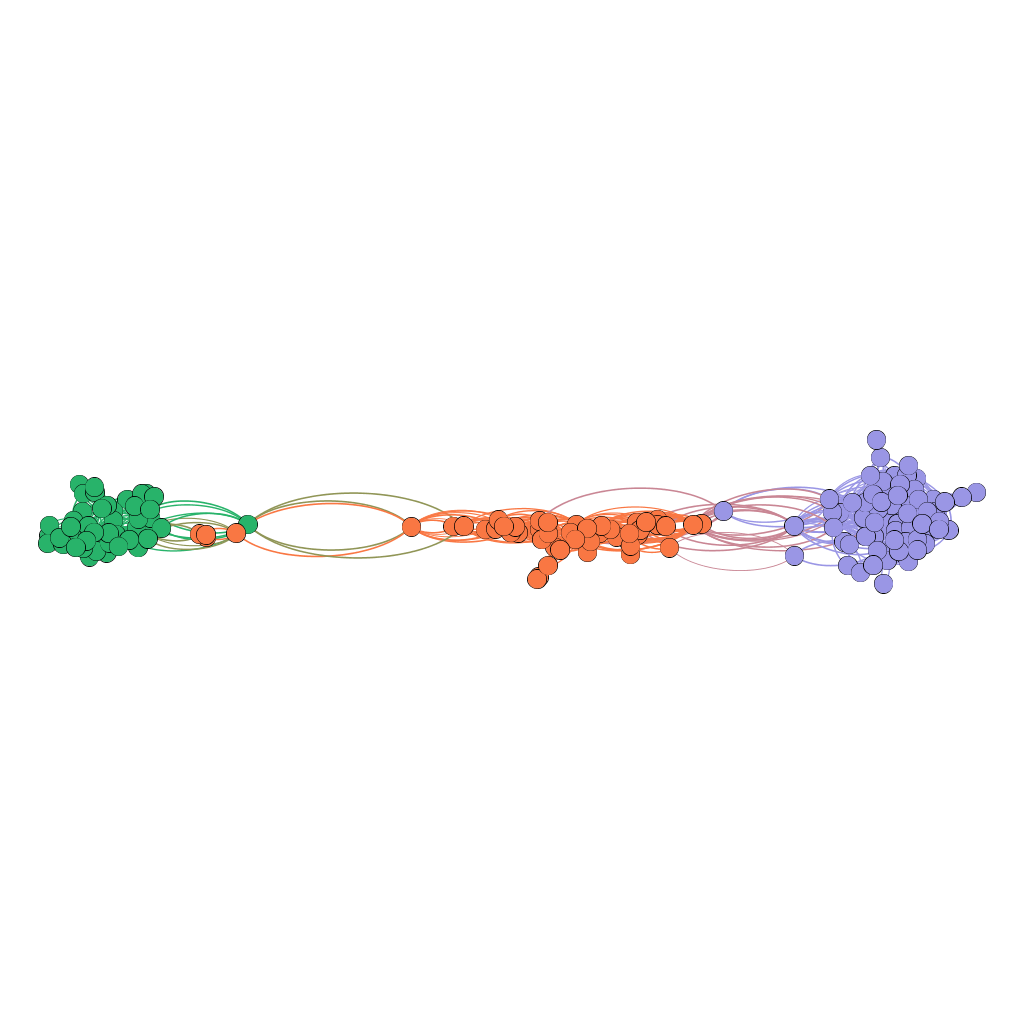
The Force Atlas Layout algorithm is a good way to represent network clusters. The distance between the nodes represents how different the nodes are to each other. The closely related nodes are bunched together and form a community as shown in Fig. 9. The node color represents each of the Cultivars. Green as Cultivar 1, Orange as Cultivar 2 and Purple as Cultivar 3. This helps in visualizing edges and how each of the communities have related to each other. Force Atlas layout also highlights the 3 outlier orange nodes that represent the Cultivars 2 that are closely related to the Green or Cultivars 1 community. All the 3 nodes are Type 1 errors of False Positive, they are incorrectly classified as Cultivar 1 community instead of Cultivar 2.

Figure 9. Force Atlas Layout of wine dataset.

# V. Conclusions

Gephi and Orange provide force-based algorithms to improve graph readability and optimization techniques to analyze geographical data (Delta Airlines), social data (Twitter), and tabular data (Wine). GeoLayout was able to position each airport according to its longitude and latitude coordinates to display the differing modules and connected airports in a comprehensive way. The nodes were ranked according to their betweenness centrality, and this articulated the reliance of the Delta operation on Atlanta as a major hub, which could lend to its potential fragility as an airline. Twitter analysis was able to identify different communities in the social network, which revealed public opinions concerning coronavirus through a metadata tag, #hashtag. Similar to political tension, the hashtag structure shows significant signs of polarizing views concerning public health, including politically charged hashtags. One of the most defining features of anti-vaccine communities is the intra-node interactivity, which suggests that the community acts more as an “echo chamber” and subsequently does not have a diverse set of outbound edges to other communities. Network classification helped classify the wine dataset into three communities that represent the three different cultivars. Visualization algorithms grouped similar wines closer to each other and the interdependencies between the cultivars. Additionally, this helped in pointing out 3 nodes with Type 1 errors of network misclassification.

# References

[1] *What is graph analytics, and why is it important?* Artificial Intelligence, Big Data Analytics and Insight.

(2021, February 24). Retrieved December 18, 2021, from https://www.analyticsinsight.net/graph-analytics-important/

[2] *The Open Graph Viz Platform*. graph exploration and manipulation. (n.d.). Retrieved December 18, 2021,

from https://gephi.org/

[3] Bioinformatics Laboratory, U. of L. (n.d.). *Data Mining*. Orange Data Mining - Data Mining. Retrieved

December 18, 2021, from https://orangedatamining.com/

[4] *GeoLayout Plugin*. Gephi. (n.d.). Retrieved December 18, 2021, from

https://gephi.org/plugins/#/plugin/geolayout-plugin

[5] Jacomy, M., Venturini, T., Heymann, S., & Bastian, M. (n.d.). *FORCEATLAS2, a continuous graph layout*

*algorithm for handy network visualization designed for the Gephi software*. PLOS ONE. Retrieved

December 18, 2021, from https://journals.plos.org/plosone/article?id=10.1371%2Fjournal.pone.0098679

[6] Fruchterman, T. M. J., & Reingold, E. M. (2006, October 30). *Graph drawing by force‐directed placement*.

Wiley Online Library. Retrieved December 18, 2021, from https://onlinelibrary.wiley.com/doi/10.1002/spe.4380211102

[7] *Circle Packing Methods*. Observable. (n.d.). Retrieved December 18, 2021, from

https://observablehq.com/@mbostock/circle-packing-methods

[8] *Analyzing the connectedness of an airline's routes*. (n.d.). Retrieved December 18, 2021, from

https://studentwork.prattsi.org/blog/2018/11/05/analyzing-the-connectedness-of-an-airlines-routes/

[9] Partow, A. (n.d.). *The Global Airport Database*. . Retrieved December 18, 2021, from

https://www.partow.net/miscellaneous/airportdatabase/

[10] Curran, A., Hardiman, J., Singh, J., & Pande, P. (2020, August 16). *Hub and spoke vs point to point - how*

*COVID will change both models*. Simple Flying. Retrieved December 18, 2021, from https://simpleflying.com/hub-and-spoke-vs-point-to-point/

[11] J. Frankenfield, “What is social data?,” *Investopedia*, 19-May-2021. [Online]. Available:

https://www.investopedia.com/terms/s/social-data.asp. [Accessed: 18-Dec-2021].

[12] UCI Machine Learning Repository Wine Data Set from Stefan Aeberhard July 1, 1991 from

<https://archive.ics.uci.edu/ml/datasets/Wine>