Hobbies Network

Enrico Romano
Simone Lu
Leonardo Crescenzi

TEAM WORK

Enrico Romano

Leonardo Crescenzi

Simone Lu

Index

Dataset Description

- Our Dataset analyze Slovakian people's hobbies and interests, aged between 15-30.
- Data have been collected in 2013 from statistics students in FSEV UK who invited their friends to participate in the survey.
- The dataset consists of 1010 rows and 150 columns (139 integer and 11 categorical)

Scan For The Dataset

Data Cleaning

1. We started by choosing 30 variables referring to music and film genres.
2. We have build a correlation matrix in order to understand the correlation between our selected variables.
3. We have created a new dataset with 3 columns: var1, referring to the source of the network; var2, referring to the target of the network; value, referring to the correlation between variables, the weight of each connection

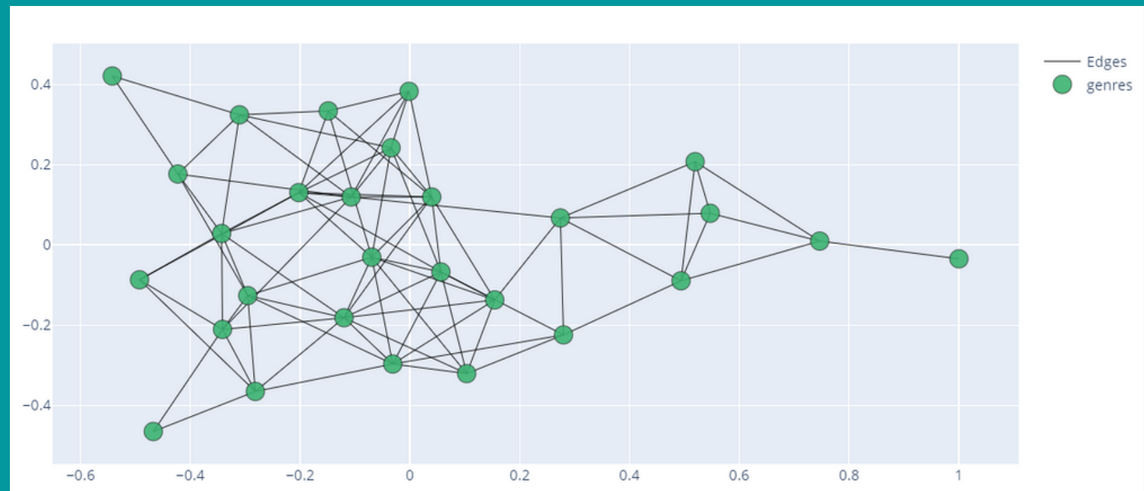
Objective

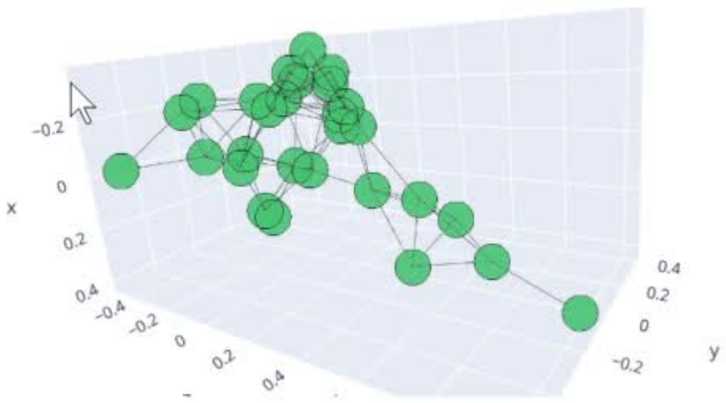
Find clusters that share similar hobbies in order to perform marketing analysis

Network Representation

What is the graphical representation of our network?

1. We represent our network with the Kamada Kawaii layout.
2. From the plot in the right side, we can give a first look at the interactions inside our network.
3. For example, we can notice that if someone is interested in Fantasy, probably is interested also in Comedy, Folk and Romantic.





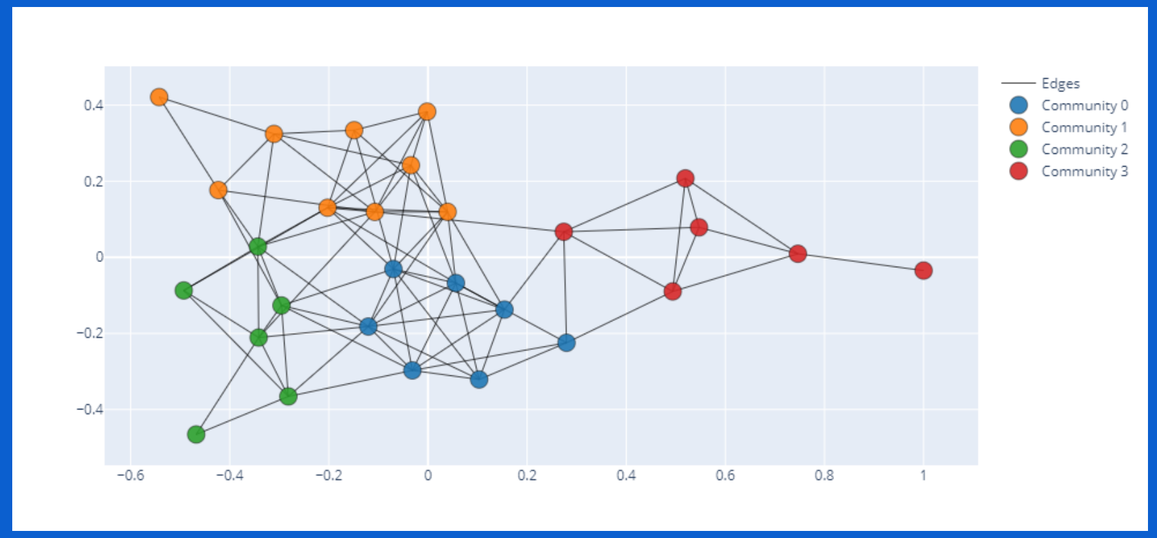
Network Communities

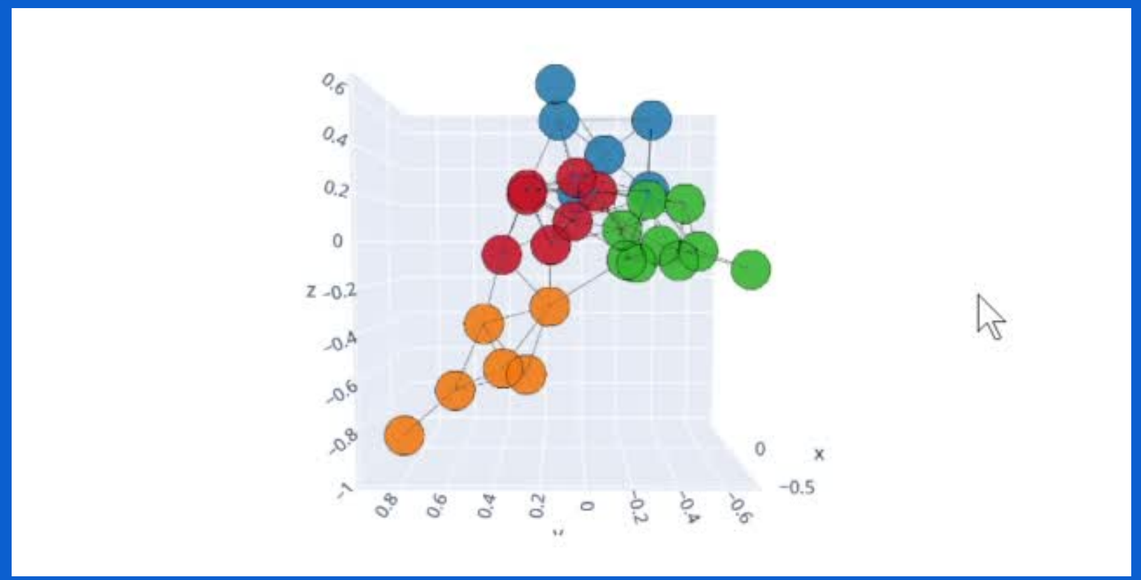
How many communities are in the network? What insight can we extract from out communities?

In order to compute the community detection, we have performed both Greedy and Louvain algorithm.

We noticed that with Greedy modularity, we have a sub-optimal partition because some nodes were assigned to different communities So, we preferred Louvain.

We obtained four different clusters, as we can see in the figure on the left.



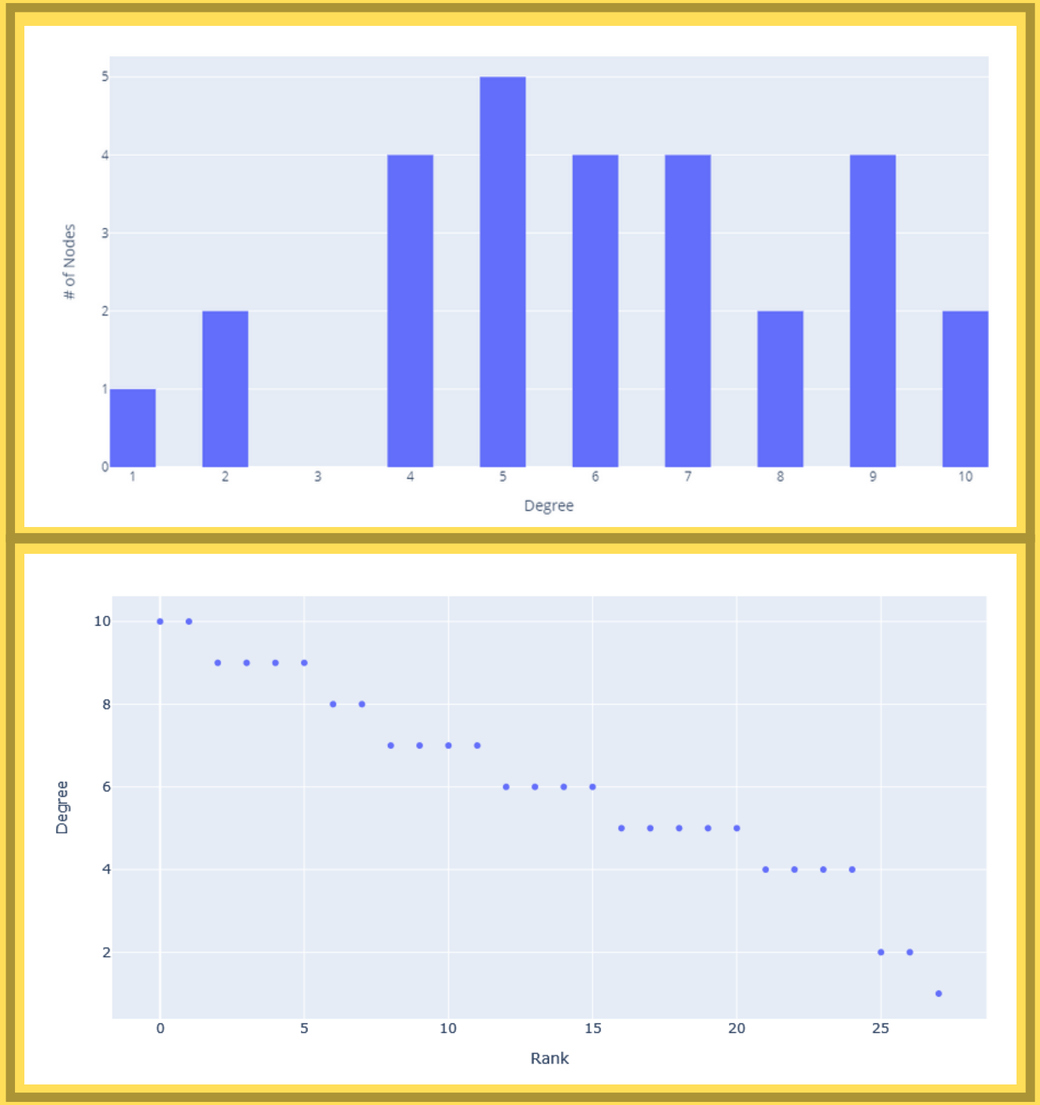


Network Degree

What is the highest number of degree?

Thanks to the degree histogram, we can notice that the most common degree is 5. Moreover, with the degree rank plot, we can also notice that five nodes have five degrees

Indexes



Network Analysis

We focused our network analysis on a few simple questions, which helped us to better understand its structure

Results

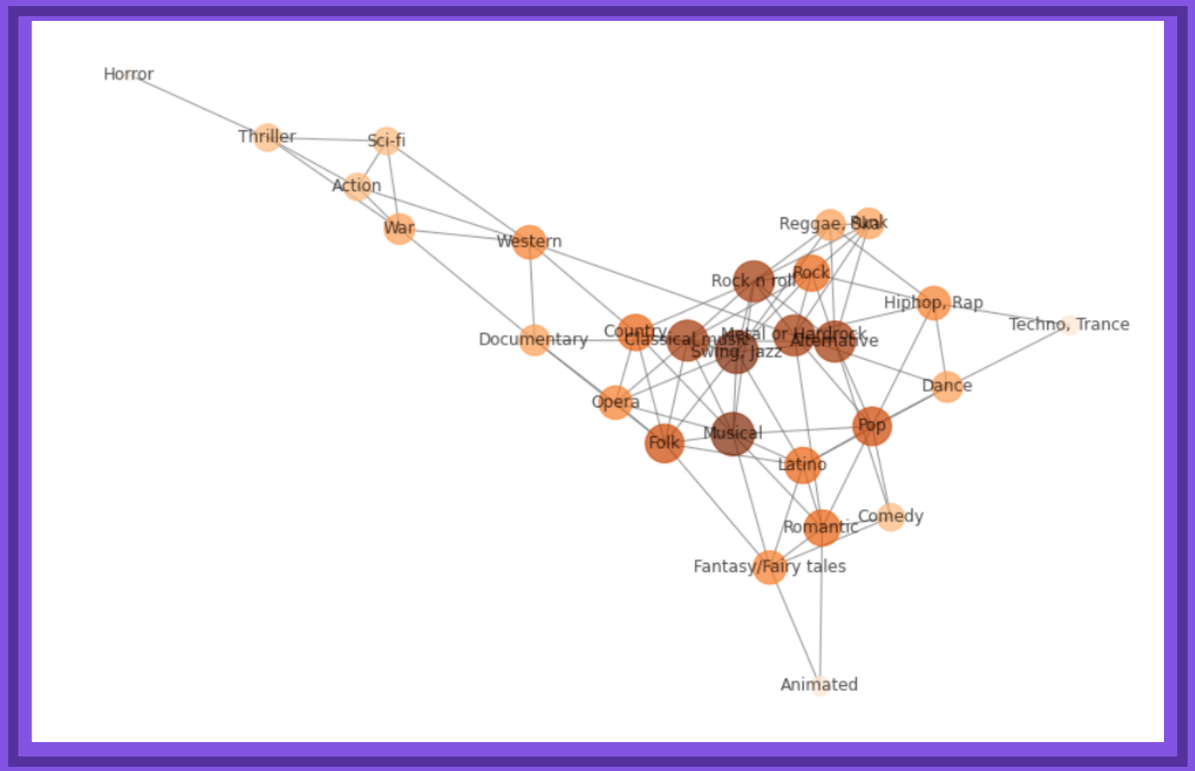
```
Is the graph connected? True
Is the graph bipartite? False
What is the diameter of the network? 6
Is the graph directed? False
What is the average degree of the nodes? 6.071428571428571
```

Network Analysis

Degree Centrality is simply : Degree of a node - number of edges

Thanks to the scale of orange, we ca notice that"Musical" and "Swing, Jazz" have the highest degree, "Horror", instead, has the lowest degree

Degree Centrality

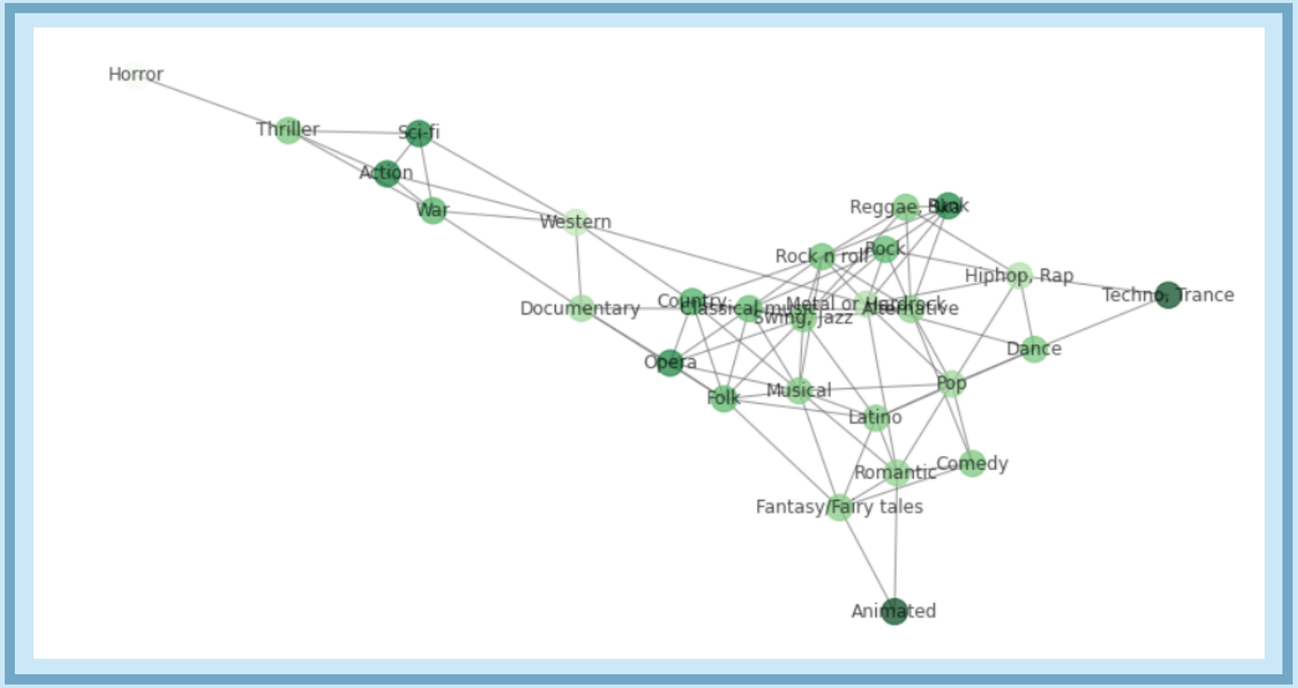


Network Analysis

Nodes' Clustering Coefficient

Clustering coefficient of node, is defined as the likelihood that any two nodes with a common neighbour are themselves connected

Horror has the lowest coefficient.
Techno, the highest.

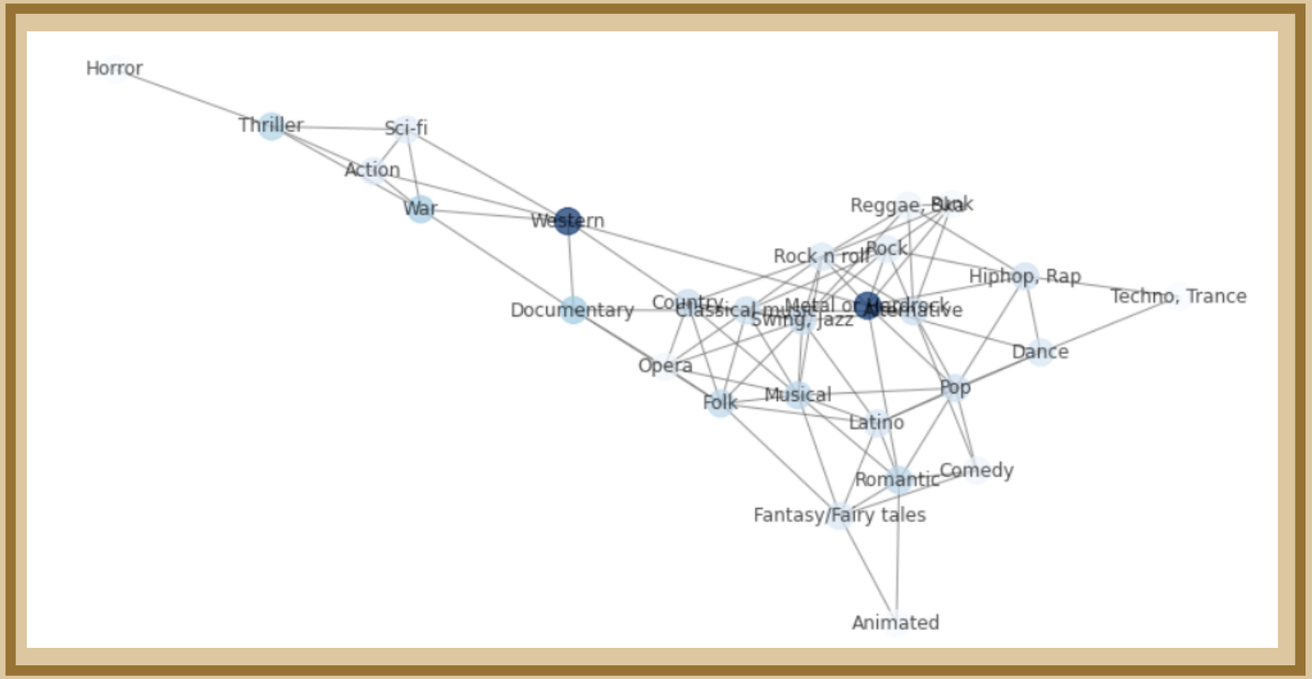


Network Analysis

Betweenness centrality of a node v, is the sum of the fraction of all-pairs shortest paths that pass through v

In this case, 'Metal' and 'Western' have the highest betweenness, while 'Horror', the lowest

Betwenness



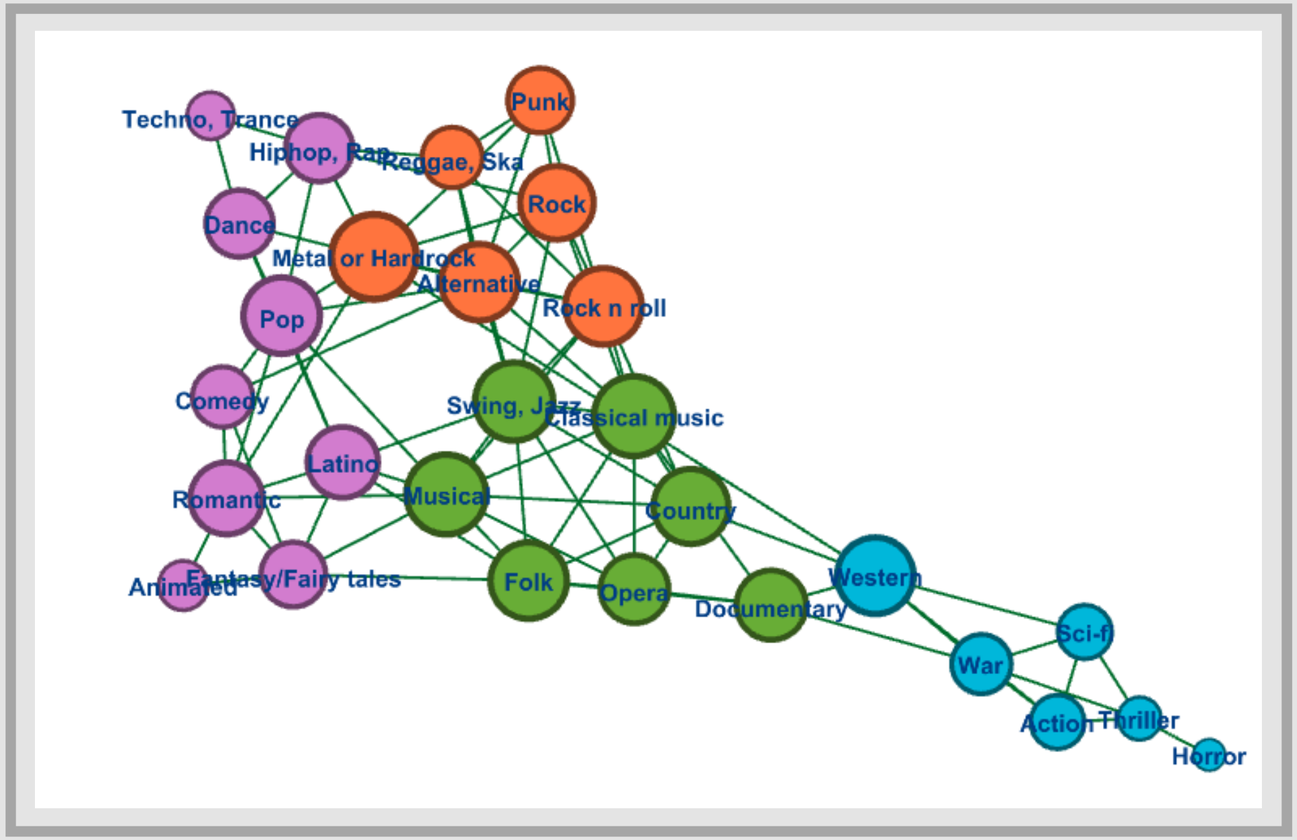
Using Gephi

Layout: Force Atlas

Color: Modularity Class

Size of nodes: Closeness Centrality

Community Detection



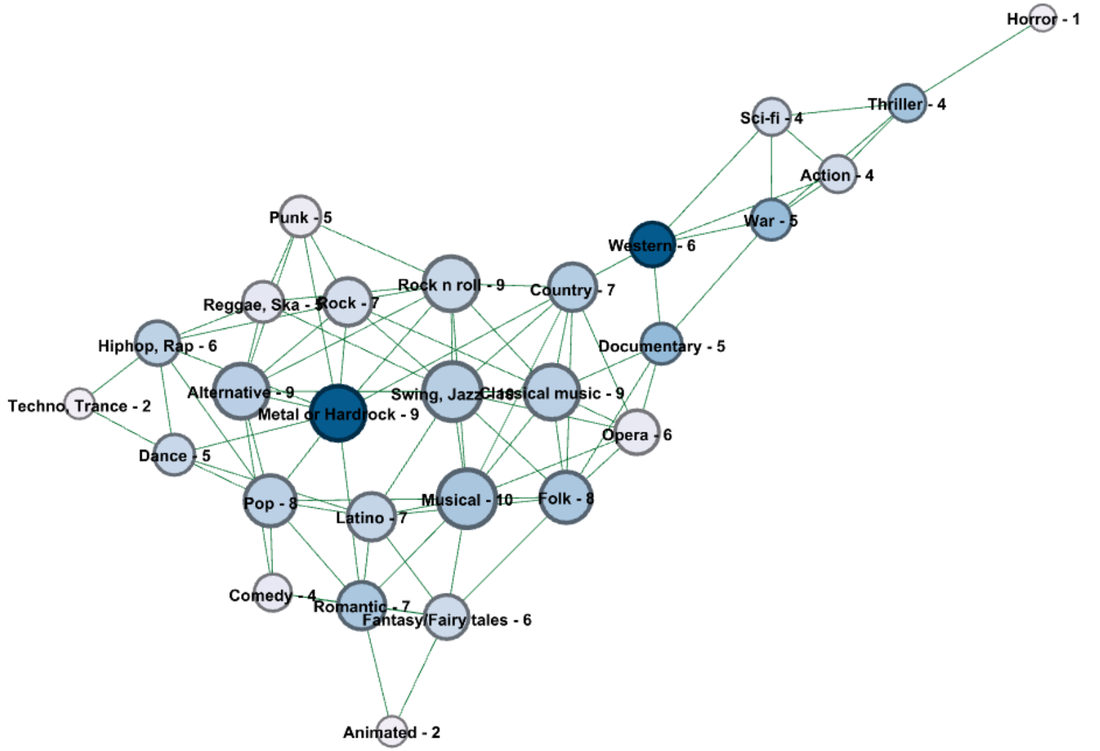
Using Gephi

Layout: Yifan Hu

Color gradient:
Betweenness centrality

Size of nodes: Degree

Nodes Importance



Conclusion

Community 1
Country, Opera, Folk, Classical, swing/jazz, Documentary, Musical

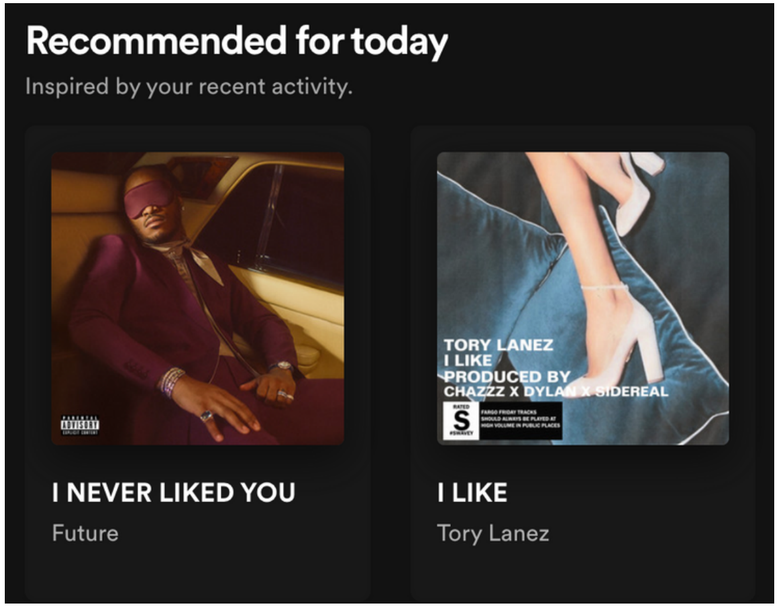
Community 2
Sci-fi, Western,Action, Thriller, War,Horror

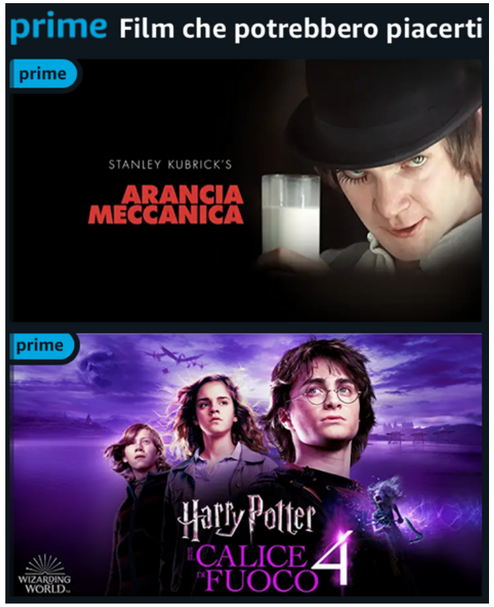
Community 3
Metal or Hardrock,Rock, Rock n roll, Alternative, Techno, Trance, Hiphop, Rap, Dance,Punk,Reggae, Ska

Community 4
Musical, Fantasy/Fairy tales, Latino, Romantic, Comedy, Animated, Pop

Possible Scenarios

This kind of network analysis works like a recommendation system and allow future development like suggestions on music and films based on choices made by any user.





THANKS FOR YOUR ATTENTION!