Music recommendation network analysis: algorithm-generated versus human-generated

Pietro Ghedini - 0001087345
pietro.ghedini3@studio.unibo.it

Abstract

*Since from the beginning, music always generated communities, namely communities related to artists belonging to the same genre class but also communities of listeners, allowing people to connect.
In this work, I consider a network, constructed using data obtained from Spotify's web API service. The nodes are the artists and a link appears for each Spotify's suggestion. I analyze different centrality measures and the relation between the number of followers and the popularity of the artists, trying to find some analogies between the obtained results.
I then compare this algorithm-generated network with the one obtained from AllMusic database, which instead is human-generated. In this way, it's possible to check if the Spotify's algorithms are in agreement or in disagreement with the human perception of music.*

Introduction

The diffusion of streaming platforms is a very important thing for the modern music industry, since it makes music easy to access and portable. In addition, it influence industries and labels in their decisions on which artists or genres they should invest and on which not. Globally, Spotify is one of the most used (if not the most used) streaming platform for music, with more than 551 million users [1].
The idea of this project came from the usage of Spotify as a music streaming platform, where, each time you open the profile of an artist, this is what you see:

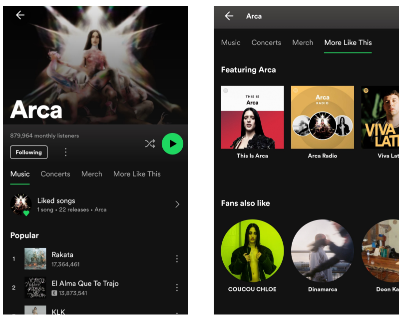

Figure 1: *Example of what appears in the Spotify app when you open an artist page and when you click on more like this.*

The starting point was to analyze in details the Spotify network, since we can get more informations about the single artists using Spotify API[1]. After analyzing the topology of this network, different centrality measures were considered to capture different characteristics of the network. In this way is possible to identify and quantify the importance of a node in a network from different points of view: in particular, degree centrality (also in and out), closeness centrality, eigenvector centrality and betweenness centrality are taken into account. A community identification algorithm is also used.
After having studied in detail the Spotify network, I decided to compare it to the network obtained from AllMusic[2] as these two streaming platforms follow two different recommendation criterion: for the first one, suggested artists are determined by algorithms based on users activity; the second one, instead, is based on the judgement of an editorial group, made up of a substantial number of music experts, even though users contributions are accepted.
In order to compare them, I had to recreate the Spotify network since not all of the starting nodes are in the AllMusic database. All of the previous computations are repeated for the new networks and then the results are compared to find if is possible to find analogies or if human-generated and algorithm-generated networks are different.

Methodologies

A network (or graph) is a collection of nodes (or vertices) connected via links (or edges), is a simplified representation that reduces the system to an abstract structure, capturing only the basics of connection patterns and little else. To capture more details of the system, we can label edges and nodes [2].
The data were obtained differently for the two streaming platforms: for Spotify, I used web calls through their API service, where you put the URI of the artists

[1] https://developer.spotify.com/

[2] https://www.allmusic.com/

and you obtain the informations that you need [3]; for AllMusic, I used the research bar and copied the names of the artists that where recommended for the one I was looking for.
The Spotify dataset is made up starting from 100 artists (the one in yellow in the table that can be found in the GitHub repository [4]) that I listen to and seeing the artists that Spotify suggests for each one of them: all these names represent the nodes of the network. An edge is added everytime the suggestion is present: in particular, Spotify's system suggests 20 other artists for each one. Since there are cases in which artist B appears as related to artist A but not viceversa, the network that is obtained is a directed network, meaning that the graph is asymmetric as the edges have a specific direction (from node A to node B for example).

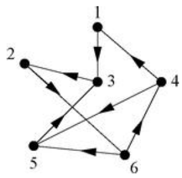

Figure 2: *Example of a directed network, where we see that for each edge there is an arrow specifying the direction. Given the same nodes and the same links, the undirected network version of this would be without the arrows on the edges [2].*

As a first thing, I created the adjacency and the incidence matrices for the network and I made a plot of the popularity (value from 0 to 100 attributed to each artist by Spotify's algorithm based mostly on users activity) and the number of followers.
I considered then different centrality measures, to find the most important nodes in a network from different points of view. Since the concept of importance can be defined in different ways, there are different measures, capturing different aspects.
Firstly, I considered the degree centrality, or degree, which is the simplest measure. The degree k_i of a vertex i is the number of connections of that vertex: for a direct graph, we can distinguish the in-degree (number of ingoing links in a node) from the out-degree (number of outgoing links from a node) [2]. As an additional information, the average degree of the network was computed. To characterize the network I also computed the average clustering coefficient: the clustering coefficient is a measure of the transitivity in a network, defined as the probability that two network neighbours of a vertex are also neighbours of each other [2] (so if artist A is similar to artist B and artist C, also B and C are similar [3]).
Using the values of the different degree-measures, is possible to plot the correspondent cumulative degree distributions $P(k)$, whose shape can help to identify the type of the network. Networks with $P(k)$ following a power law ($P(k) \sim Ck^{-\alpha}$) are also called scale-free networks: this is a characteristic of real networks, differentiating them from random ones, which have a Poisson distribution [4].
Continuing with the centrality measures, I considered the betweenness centrality: this is defined as the ratio between the number of shortest paths passing through a node and the total number of shortest paths in the network, measuring the extent to which a vertex lies on the paths between other vertices.
Closeness centrality measures the mean distance from a vertex to the other vertices: this measure is related to the ability to communicate with the other nodes, meaning that nodes that are closer to the others are more important.
Eigenvector centrality gives a score to each vertex proportional to the sum of scores of the neighbours. This is the most solid centrality measure since it has a very grounded mathematical background and it is useful to examine community structure.
In conclusion, the last centrality measure used to study the network is a variation of the eigenvector centrality, the so called PageRank measure, that gives back a ranking of the nodes based on the structure of the incoming links [2].
The results to all these measures will be presented in the following section. In addition, I created also plots involving the different results for the centrality measures considered and the Spotify parameters, which are popularity and the number of followers to see if there is correlation between them.
Lastly, the $greedy - modularity$ algorithm was used to find the different communities in the network: this algorithm is greedy in the sense that chooses the best possible option at each step, without considering the consequences on future steps, as it works maximizing modularity, a measure of the extent to which like is connected to like in a network [2], at each step.
After having studied in details the network obtained from Spotify, following the idea behind [3], I decided to compare this graph, which is algorithm-generated, with the AllMusic one, human-generated, where the related artists are chosen by a group of musical experts, accepting also contributions from users. In order to do this comparison, I had to remove some of the initial 100 nodes from the Spotify network, going down to 82, as not all of the initial artists are in the AllMusic database. This allows anyway also to make a comparison between the initial and the reduced Spotify network.

[3] It's important to underline that the Spotify data were taken during the first week of June 2023 and they are updated after some time, so now the results could be different.
[4] `https://github.com/pietroghedini/Complex-Network-exam-project.git`

All the measures indicated above are re-evaluated for the two new networks.

Results

The Python code used to obtain all the results presented in this section, with the correspondent data files can be found in the GitHub repository.

Spotify network

The network built starting from the Spotify data for the 100 artists chosen is the following. This directed graph is made up of 1028 nodes and 2000 edges.

Spotify network

Type	Nn	Ne	<k>	C
Directed	1028	2000	3.89	0.12

Table 1: *Summary of the network parameters: Nn is the number of nodes, Ne is the number of edges, $< k >$ is the average degree and C is the average clustering coefficient.*

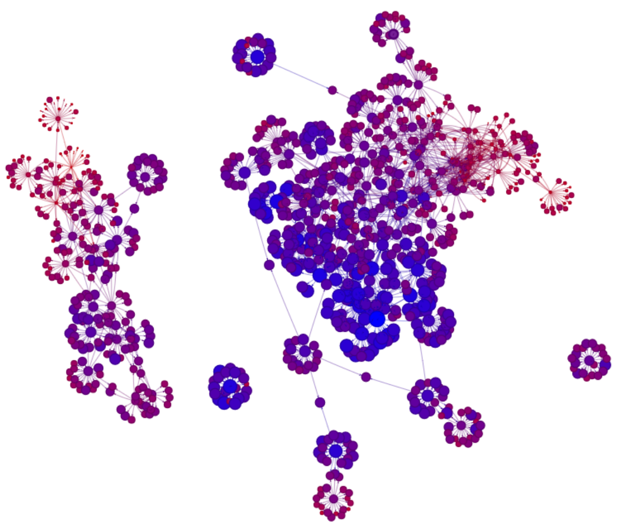

Figure 3: *Spotify's network: the nodes are coloured and have a size based on the popularity value (from 0=red to 100=blue) attributed by Spotify.*

Looking at Figure 3 it's possible to see that there are 4 different components in this network: the biggest one includes the majority of non-Italian-singing artists, the second biggest one is instead made up of all the Italian-singing artists and there are two isolated nodes, so two artists that are not connected to any other artist in the network, neither are their related artists. Even inside the biggest component, it's possible to identify a denser area, where there are more links, in respect to the whole component. One can also see that this denser area is characterized by links between artists with lower popularity. It's interesting to note that the artists in the denser area are characterized by same genres, or subcategories of those genres. The adjacency and incidence matrices can be found in the GitHub repository.

By simply looking at the network, only 4 different components appear as already said, but using a community identification algorithm, like the *greedymodularity* one included in the *NetworkX* library of Python, a more complex structure can be found, as it's possible to see in Figure 4.

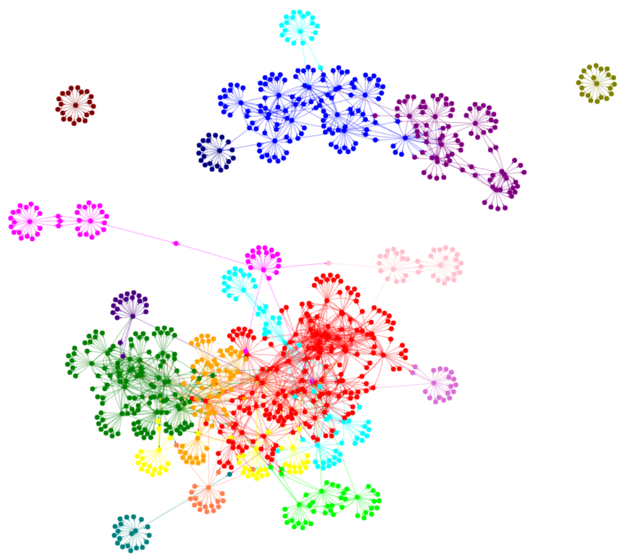

Figure 4: *Spotify's network where the different communities, 18 in total, are highlighted in different colours.*

Inside the two biggest components there is a division. For the Italian-singing artists, the algorithm identifies 4 different communities, while the non-Italian-singing artists there are 12 different components. In both cases there are some communities bigger than the others. By looking at the nodes in the different communities, is possible to find some relations between the artists based on the genre/genres they play, even though this is not always the case.

Let's consider the Italian-singing group for example: there is one node (dark blue) which is isolated from the others and it represents the artist Meg: if we look at the genres she plays, there are some which are in common with other Italian artists but the fact is that most of the artists she's related to are Asian people. In the two biggest communities (blue and light purple) there is a distinction of the artists between the indie/alternative/experimental scene of the Italian music and the class of the "cantautori" and pop Italian singers.

A similar analysis can be also done with the other component of the network, where also there the communities found share the same (or derivation of the main) genres: in particular the red component, the biggest one, includes the majority of the elec-

tronic/experimental/hyperpop artists, related also by the fact that they are under the same labels and there are different collaborations between most of them. Another interesting case is the one of the "navy" colour node, which represents the Spanish-singing artists, mainly related to the reggaeton (and subgenres) scene.
It's interesting now to plot the number of followers of the artists as a function of the popularity value: lower value of popularity correspond, in the majority of cases, to lower number of followers. It's also possible to see a consistent number of high popularity artists but with a small number of followers and a mismatch between the most popular one according to Spotify (Taylor Swift) and the one with highest number of followers (Ariana Grande). It's more difficult to say something about the low popularity artists since there are four of them with value equal to one: what I can say is that the lowest value for the number of followers (51 in this case) corresponds to an artist with popularity equal to 1 (Sierra Moreno). The same plot was made also in logarithmic scale in order to find the type of relation that exists between the two parameters.

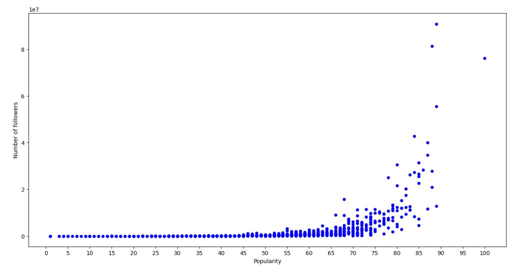

Figure 5: *Number of followers and popularity plot.*

A log-log plot allows to better identify the relation between these two parameters.

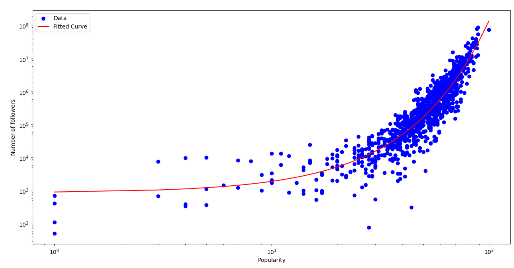

Figure 6: *Number of followers and popularity plot in logarithmic scale.*

What is found is that the underlying relation is of exponential type $y = A * 10^{b*x} + C$, where x is the popularity and y the number of followers. Using curve fitting algorithm, the best-fit curve found has equation $y = (2.3 * 10^{-2}) * 10^{1.2x} + 2.9$.
Considering the different centrality measures, I decided to show and compare just the first five artists, which will represents the five most important artists among all as captured by the different measures, and the last five artists, representing the less important ones. I also add a ranking based on the popularity and number of follower. The results are shown in Table 2 and in Table 3.

Popularity	Followers	Degree
Taylor Swift (100)	Ariana Grande (90925787)	Namasenda (44)
SZA (89)	Billie Eilish (81430807)	A.G. Cook (41)
Rihanna (89)	Taylor Swift (76134372)	SOPHIE (41)
Ariana Grande (89)	Rihanna (55600948)	Hannah Diamond (38)
Miley Cyrus (88)	Selena Gomez (42790137)	GFOTY (38)
⋮	⋮	⋮
POBBLES (3)	Bitch Volley (337)	Portishead (1)
Sierra Moreno (1)	ZEF & MARZ (311)	Broadcast (1)
maid boys (1)	maid boys (110)	The Sugarcubes (1)
A-bee (1)	INPASTE (77)	The Knife (1)
Hierophant-Green (1)	Sierra Moreno (51)	PJ Harvey (1)

Table 2: *Summary of the centrality measures (part 1).*

Betweenness	Closeness	Eigenvector	PageRank
Charli XCX ($8.8*10^{-3}$)	SOPHIE ($3.2*10^{-2}$)	Planet 1999 (0.26)	Namasenda ($2.3*10^{-3}$)
Tove Lo ($5.4*10^{-3}$)	A. G. Cook ($3.2*10^{-2}$)	Namasenda (0.26)	A. G. Cook ($2.2*10^{-3}$)
Rina Sawayama ($4.3*10^{-3}$)	Shygirl ($3.2*10^{-2}$)	A. G. Cook (0.26)	SOPHIE ($2.2*10^{-3}$)
FKA Twigs ($3.4*10^{-3}$)	Charli XCX ($3.1*10^{-2}$)	GFOTY (0.26)	GFOTY ($2.1*10^{-3}$)
Namasenda ($2.6*10^{-3}$)	Namasenda ($3.1*10^{-2}$)	EASYFUN (0.25)	Hannah Diamond ($2.1*10^{-3}$)
⋮	⋮	⋮	⋮
Xiu Xiu (0)	Florence + The Machine (0)	Florence + The Machine ($9.8*10^{-16}$)	Christine and the Queens ($0.9*10^{-3}$)
Broadcast (0)	Tyler, The Creator (0)	Tyler, The Creator ($9.8*10^{-16}$)	Daft Punk ($0.9*10^{-3}$)
The Sugarcubes (0)	ROSALÍA (0)	ROSALÍA ($9.8*^{-16}$)	Peggy Gou ($0.9*10^{-3}$)
The Knife (0)	Flume (0)	Flume ($9.8*^{-16}$)	aya ($0.9*10^{-3}$)
PJ Harvey (0)	NAVA (0)	NAVA ($9.8*^{-16}$)	MACE ($0.9*10^{-3}$)

Table 3: *Summary of the centrality measures (part 2).*

Since the network is built using 100 starting nodes, is not surprise that in the higher positions of this different rankings there are those artists, exception made for the eigenvector centrality where the the first node, namely Planet 1999, is not one of the initial names, as well as EASYFUN.
Charli XCX appears as the most central node under the betweenness centrality point of view: this makes sense as if we look at her discography, we see collaboration with both artists belonging to the electronic-experimental genre (like SOPHIE or A. G. Cook) and to the the pop genre (like Tove Lo or Rina Sawayama), the two main genres the artists in this network belong to. This is consistent also with the genres Spotify attributes to her (artpop for example is a subgenre of pop music which incorporates sounds from electronic music among other genres).
It's interesting to see that Italian-singing artists ap-

pear only in the lowest positions of this measures (NAVA, MACE, INPASTE, ZEF & MARZ).

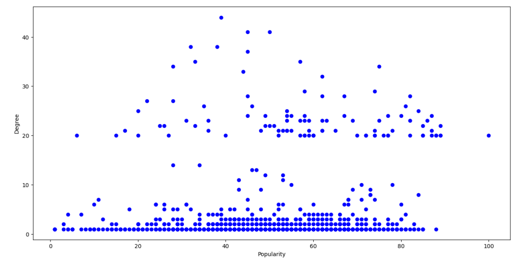

Figure 7: *Plot of the degree results as a function of the popularity value.*

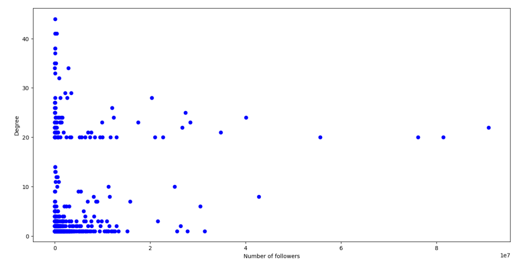

Figure 8: *Plot of the degree results as a function of the number of followers.*

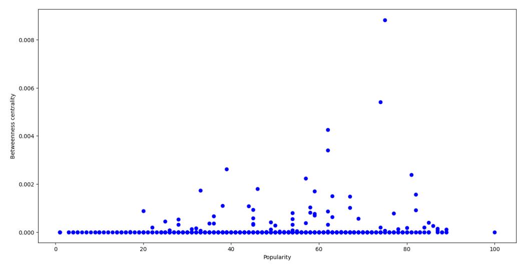

Figure 9: *Plot of the betweenness centrality results as a function of the popularity value.*

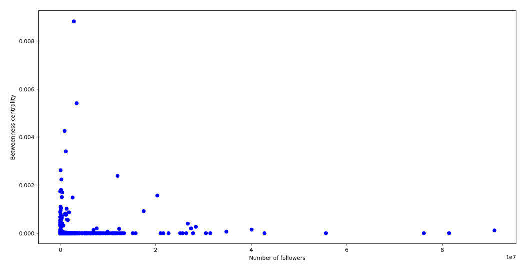

Figure 10: *Plot of the betweenness centrality results as a function of the number of followers.*

It's interesting to plot the different centrality measures as a function of the Spotify parameters, namely

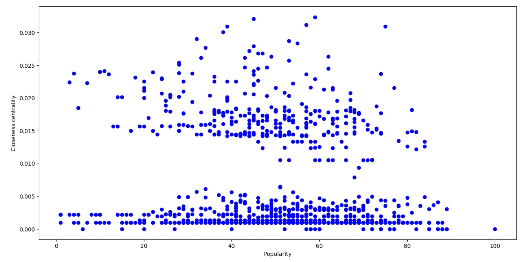

Figure 11: *Plot of the closeness centrality results as a function of the popularity value.*

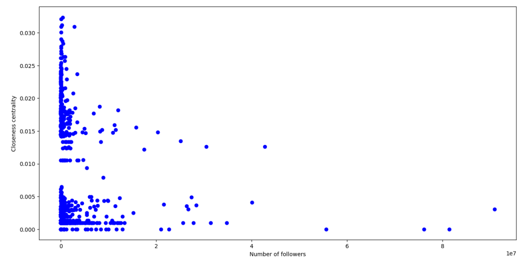

Figure 12: *Plot of the closeness centrality results as a function of the number of followers.*

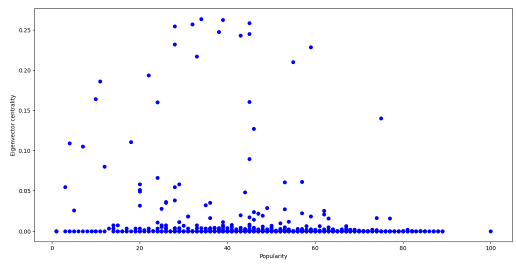

Figure 13: *Plot of the eigenvector centrality results as a function of the popularity value.*

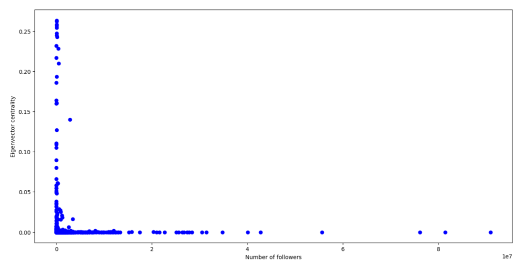

Figure 14: *Plot of the eigenvector centrality results as a function of the number of followers.*

popularity and followers, in order to find which of these measures correlates better to the parameters. The results, obtaining considering the popularity/number of followers value from smaller to bigger, are shown from Figure 7 to 16. It's evident that there is not a good correlation between the centrality mea-

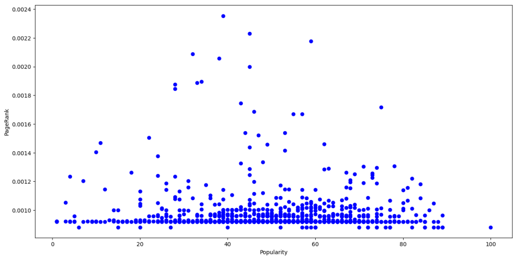

Figure 15: *Plot of the PageRank results as a function of the popularity value.*

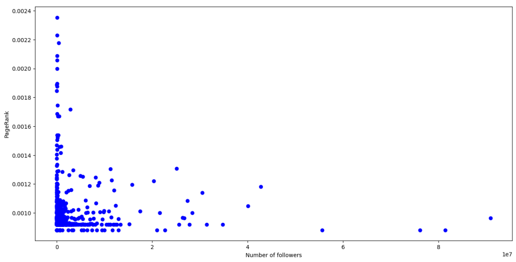

Figure 16: *Plot of the PageRank results as a function of the number of followers.*

sures and the Spotify parameters.
Focusing on the plots involving popularity, the points of the plot appear to be divided in two groups for most of the cases. Moreover, the points are equally distributed and by saying this what I mean is that there are no big differences between high/low popularity and high/low values of the centrality measures, even though the highest centrality measures are in the intermediate range for the popularity.
Regarding the plots involving the number of followers, a more regular structure appears with the artists with the lower number of followers having the highest values of the centrality measures. Also in a few of these plots, the points appear to be divided in two groups.

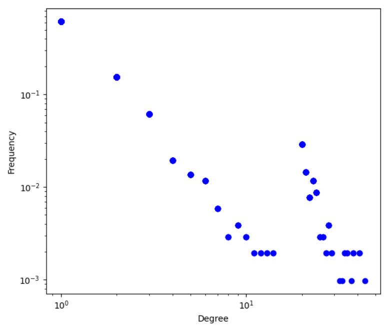

Figure 17: *Degree distribution of the network*

Lastly, I considered the cumulative degree distributions for the degree and the in-degree. For the out-degree this was not significant as the value obtained from the measure is 20 for the initial 100 nodes and 0 for all of the others. The degree distribution plot shows one peak, probably related to the fact that there are not enough data to smooth out the distribution.
For the in-degree plot, using a linear regression method, I was able to fit a curve, proving the power law behaviour of the data (the r-squared value for the plot is 0.9941). What is obtained is shown in Figure 18.

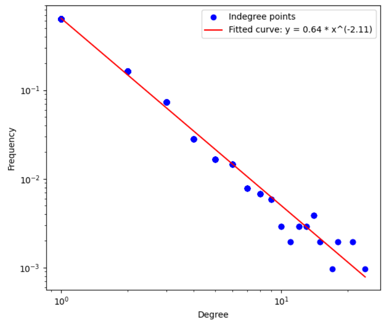

Figure 18: *In-degree distribution of the network*

AllMusic and Spotify networks comparison

After having studied Spotify's network, I decided to compare it to a similar network obtained using a different database: in this case AllMusic, since the related artists are there thanks to decisions taken by a group of music experts. So the criterion underlying the two networks is different: Spotify one is algorithm-generated while Allmusic graph is human-generated.
Not all of the 100 initial artists chosen for the previous network, but only 82, are present in AllMusic database (these artists are the one in bold among the one in yellow in the table in the GitHub repository).

Spotify network				
Type	Nn	Ne	<k>	C
Directed	823	1640	3.99	0.14
AllMusic network				
Type	Nn	Ne	<k>	C
Directed	1944	3356	3.45	0.11

Table 4: *Summary of the networks parameters: Nn is the number of nodes, Ne is the number of edges, $< k >$ is the average degree and C is the average clustering coefficient.*

First of all we see a difference by looking at the number of edges: while Spotify suggests 20 other artists for each one of them, in the AllMusic database this is not the case as one can have more or less than 20 related artists. By looking at Table 4 , it emerges that the clustering coefficient for the Spotify network is bigger than the previous case, meaning that now there are stronger ties between the nodes. Moreover, the Spotify network has also stronger bounds than the AllMusic case. The adjacency and incidence matrices can be found in the link in note 4.

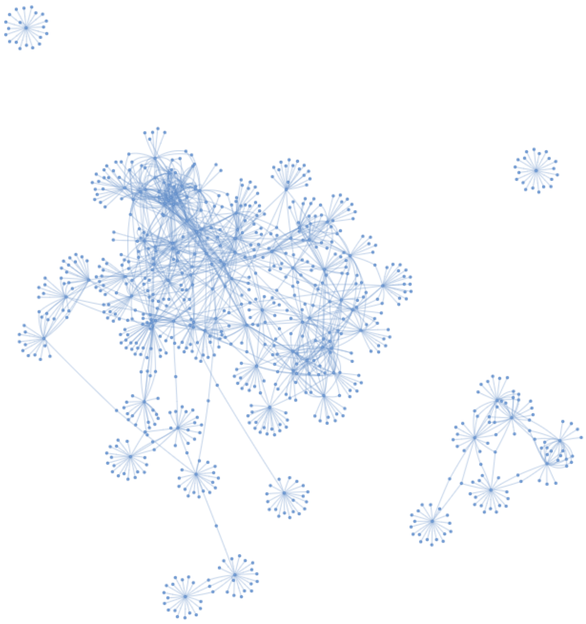

Figure 19: *Spotify's network.*

While for the Spotify network we can basically repeat the same considerations that we did in the previous section, the AllMusic network has some major differences. First of all we see 3 different components: one big component and two small components, smaller than the one appearing for Spotify. By looking at these last two, it's evident that the number of links for each node is not constant anymore. We see now no distinction between non-Italian and Italian singing artists, since now there are few connections between what are the two main components of the Spotify network (for example Elisa is now related to Dua Lipa and Florence + The Machine). The denser part of the graph now is more extended in the AllMusic case and it does not include just artists belonging to the same genres or subgenres classes.
Also now, I considered a community detection algorithm, to see what are the differences.
Comparing the Spotify result with Figure 4, the main difference is simply that some communities appear smaller and we loose one of them. The most interesting case is perhaps the AllMusic result in which we loose a distinction of genres in the communities found but also the distinction between Italian-singing and non-Italian-singing artists since now they are mixed in the same communities.

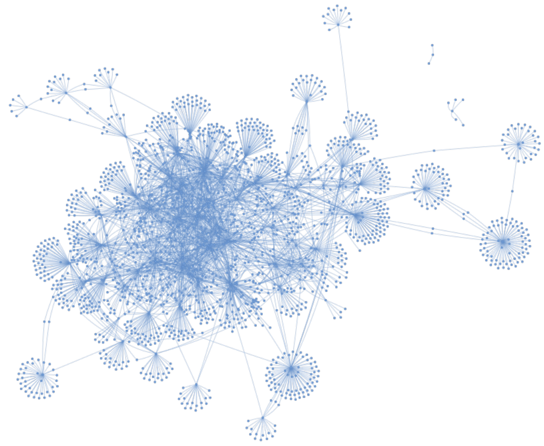

Figure 20: *AllMusic network.*

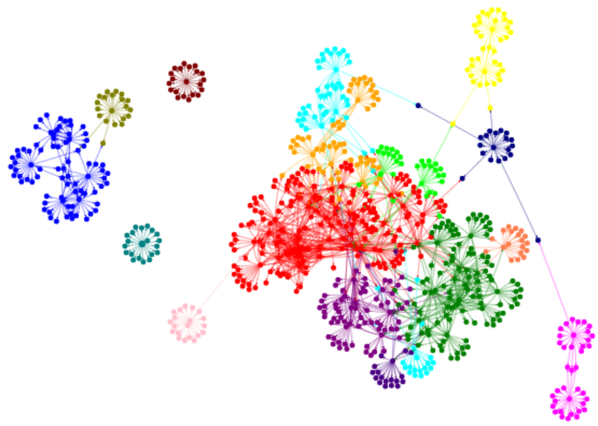

Figure 21: *Spotify's network where the different communities, 17 in total, are highlighted in different colours.*

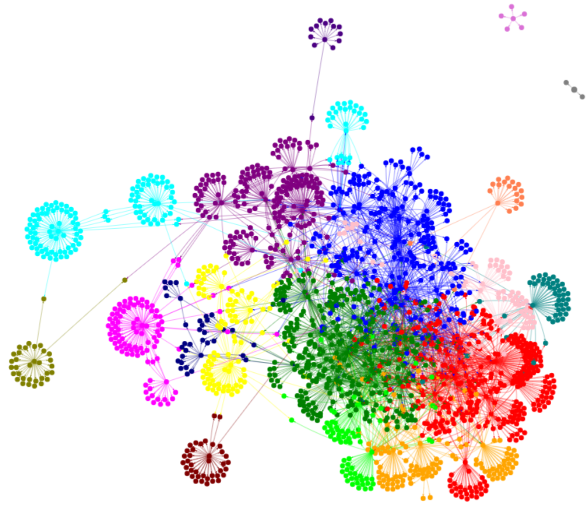

Figure 22: *Spotify's network where the different communities, 19 in total, are highlighted in different colours.*

As before, I will show the first five and the last five artists for each centrality measures: for the AllMusic case I will not consider popularity and number of followers since is an information which is not provided by the website.

Popularity	Followers	Degree
Taylor Swift (100)	Ariana Grande (90925787)	Namasenda (42)
Rihanna (89)	Billie Eilish (81430807)	SOPHIE (41)
Ariana Grande (89)	Taylor Swift (76134372)	A. G. Cook (40)
Miley Cyrus (88)	Rihanna (55600948)	Hannah Diamond (37)
Billie Eilish (88)	Dua Lipa (40089787)	GFOTY (36)
⋮	⋮	⋮
Cecile Believe (33)	umru (25276)	Leon Vynehall (1)
GFOTY (32)	daine (21806)	TDJ (1)
Hyd (28)	Hyd (18685)	CFCF (1)
felicita (28)	Vasco Brondi (15025)	Yamning Portal (1)
QT (22)	felicita (12657)	SONIKKU (1)

Table 5: *Summary of centrality measures for Spotify (part 1).*

Betweenness	Closeness	Eigenvector	PageRank
Charli XCX ($1.2*10^{-2}$)	SOPHIE ($3.8*10^{-2}$)	Planet 1999 (0.27)	Namasenda ($2.8*10^{-3}$)
Tove Lo ($0.8*10^{-2}$)	A. G. Cook ($3.8*10^{-2}$)	Namasenda (0.26)	A. G. Cook ($2.7*10^{-3}$)
Rina Sawayama ($0.6*10^{-2}$)	Charli XCX ($3.7*10^{-2}$)	A. G. Cook (0.26)	SOPHIE ($2.7*10^{-3}$)
FKA twigs ($0.5*10^{-2}$)	Shygirl ($3.7*10^{-2}$)	GFOTY (0.26)	Hannah Diamond ($2.5*10^{-3}$)
Ellie Goulding ($0.3*10^{-2}$)	Namasenda ($3.6*10^{-2}$)	EASYFUN (0.25)	GFOTY ($2.4*10^{-3}$)
⋮	⋮	⋮	⋮
Leon Vynehall (0)	Franco Battiato (0)	Franco Battiato ($1.2*10^{-15}$)	Franco Battiato ($1.1*10^{-3}$)
TDJ (0)	Sam Smith (0)	Sam Smith ($1.2*10^{-15}$)	Sam Smith ($1.1*10^{-3}$)
CFCF (0)	Christine and the Queens (0)	Christine and the Queens ($1.2*10^{-15}$)	Christine and the Queens ($1.1*10^{-3}$)
Yamning Portal (0)	Daft Punk (0)	Daft Punk ($1.2*10^{-15}$)	Daft Punk ($1.1*10^{-3}$)
SONIKKU (0)	Peggy Gou (0)	Peggy Gou ($1.2*10^{-15}$)	Peggy Gou ($1.1*10^{-3}$)

Table 6: *Summary of centrality measures for Spotify (part 2).*

Comparing the values in Table 5 and 6 with the ones of the previous Spotify network (Table 2 and Table 3), what it's possible to see is that the lowest popularity now is 22 and not anymore 1, so the artists not present in the AllMusic database are the one with a Spotify popularity smaller than 22 (for the ones I've considered). This fact is consistent also looking at the number of followers column, since now the lowest positions have an higher number than before.

Regarding the other centrality measures, the main difference is related to the lowest artists in the ranking: while before they were different for each one of the measures, now they are basically always the same for all of them. For the higher positions instead, the names are basically the same as before.

By looking at Table 7 and Table 8, it's clear the difference between the two networks (Spotify and AllMusic). The maximum degree is higher in the case of AllMusic, because the number of links per starting node is not constant and consequently because there are more nodes. Quite interesting is that considering degree we get a completely different ranking, regarding just the first five positions, in the two cases since the artists are different. Consistently with the fact

Degree	Betweenness
Lorde (158)	Charli XCX ($7.8*10^{-3}$)
Taylor Swift (119)	Carly Rae Jepsen ($5.2*10^{-3}$)
Charli XCX (112)	Halsey ($4.0*10^{-3}$)
Halsey (106)	Lady Gaga ($3.5*10^{-3}$)
Carly Rae Jepsen (106)	Lorde ($3.2*10^{-3}$)
⋮	⋮
Sofia Kourtesis (1)	Sofia Kourtesis (0)
Elite Gymnastics (1)	Elite Gymnastics (0)
CFCF (1)	CFCF (0)
Jam City (1)	Jam City (0)
Porter Robinson (1)	Porter Robinson (0)

Table 7: *Summary of centrality measures for AllMusic (part 1).*

Closeness	Eigenvector	PageRank
Charli XCX ($1.5*10^{-2}$)	Carly Rae Jepsen (0.20)	Hannah Diamond ($9.4*10^{-4}$)
Carly Rae Jepsen ($1.5*10^{-2}$)	Charli XCX (0.16)	Alice Glass ($9.3*10^{-4}$)
Lana Del Rey ($1.4*10^{-2}$)	Tove Lo (0.15)	Kero Kero Bonito ($9.0*10^{-4}$)
Lorde ($1.4*10^{-2}$)	Selena Gomez (0.14)	Hyd ($8.9*10^{-4}$)
Grimes ($1.4*10^{-2}$)	Lady Gaga (0.14)	SOPHIE ($8.2*10^{-4}$)
⋮	⋮	⋮
Loredana Bertè (0)	Loredana Bertè ($1.3*10^{-25}$)	Loredana Bertè ($5.1*10^{-4}$)
Big Red Machine (0)	Big Red Machine ($1.3*10^{-25}$)	Big Red Machine ($5.1*10^{-4}$)
Mia Martini (0)	Mia Martini ($1.3*10^{-25}$)	Mia Martini ($5.1*10^{-4}$)
Franco Battiato (0)	Franco Battiato ($1.3*10^{-25}$)	Franco Battiato ($5.1*10^{-4}$)
Daft Punk (0)	Daft Punk ($1.3*10^{-25}$)	Daft Punk ($5.1*10^{-4}$)

Table 8: *Summary of centrality measures for AllMusic (part 2).*

that the network is built starting from a set of initial nodes, the names appearing for the degree are part of this group of vertices. Looking at the betweenness column in the tables, the only common name is Charli XCX: since she's first in both of networks, this could be a hint that she's quite important and central in the network of artists considered. Considering now closeness, common names are present both in the upper (Charli XCX) and in the lower (Franco Battiato and Daft Punk) positions, even tough they are found in different positions. From the eigenvector centrality and PageRank points of view, there are some similarities in the names appearing but none of them is in the same position in both rankings.

Lastly, I considered also for this two networks the degree and in-degree distributions. Once again for the Spotify network the out-degree distribution was meaningless, while for the AllMusic case it was possible to plot it but since I couldn't compare it with the Spotify one, I neglected it. There is still a peak in the degree distribution for the Spotify case, which, even if smaller, appears also in the AllMusic case.

As before, I plotted a curve using the linear regression method. The Spotify network still follows a power law (the r-squared value is 0.9942), while for

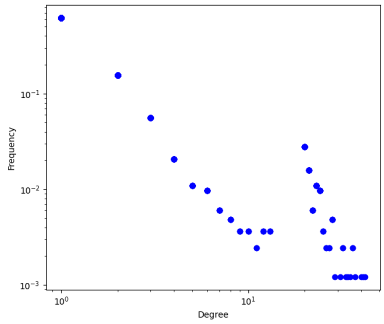

Figure 23: *Degree distribution of the Spotify network.*

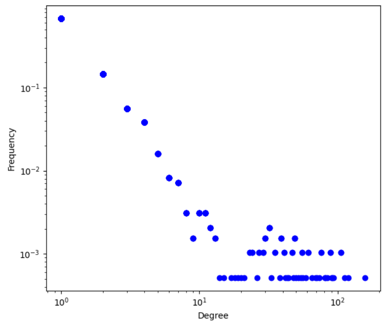

Figure 24: *Degree distribution of the AllMusic network.*

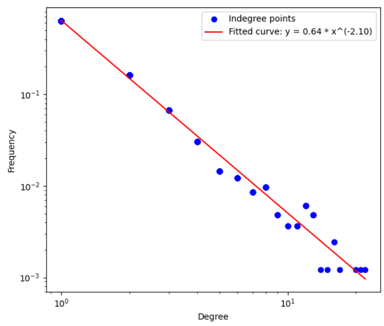

Figure 25: *In-degree distribution of the Spotify network.*

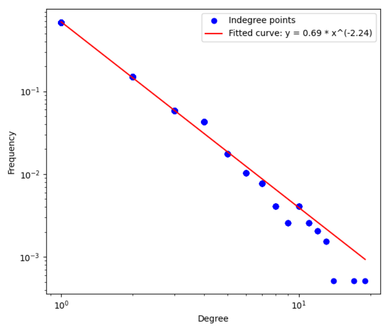

Figure 26: *In-degree distribution of the AllMusicnetwork.*

the AllMusic graph the situation is interesting: from a graphical point view, it's possible to see that the end of the curve does not follow correctly the behaviour of the data but the r-squared value is higher than the Spotify case, being equal to 0.9944, probably due to the fact that the initial part is very well approximated by this plot. The fact that the tail is not in agreement with the data could lead to think that the behaviour of this distribution is not a power law but that the obtained result is simply due to a small number of points.

Conclusions

In this work I studied a music recommendation network built using Spotify and compared it to another one built using a different database (AllMusic) chosen since the suggestions are not based on algorithms like in the first case but are made following the decisions of a group of music experts and suggestions made by users. In particular, different centrality measures where taken into account in order to try to understand how the mechanism behind music recommendation systems works. Even though the number of nodes considered is quite small, I was able to gain some key informations, starting from a big difference in how people perceive similarity between different artists and how an algorithm based on users activity, songs in common playlist and usage of the streaming platform perceive it. This is evident in the comparison between the Spotify and the AllMusic network: while in the first case all the Italian-singing artists are completely disconnected from the non-Italian-singing artists, in the ladder there are some links between the two groups (if you search the artist Elisa, you find Dua Lipa as a suggestion for example). Out of this, one could be drawn to think that Italian-singing artists will never appear as suggestions of non-Italian-singing artists in the Spotify case, but this may not be the case: a collaboration between singers of this two groups could lead to a modification of the algorithm, or more simply the initial 100 nodes chosen don't have this kind of connections but considering other artists the links could appear. If even increasing the number of singers considered the two blocks remain separated, this could be a hint of a limitation of the recommendation network used by Spotify, since it does not allow connection between specific classes of singers and so could be implemented taking into account also suggestions from users or the action of music experts.
Having a first network for Spotify and a second one with some nodes removed allow to make a comparison between the two: nothing really changes but from the clustering coefficient it emerges that the links that were removed were the weaker ones, mak-

ing the nodes more strongly tied.
An interesting conclusion is that there is not a high correlation between the results of the centrality measures taken into account and the Spotify parameters (popularity and number of followers), while there is correlation between the Spotify parameters, as seen in Figure 6. This could be a hint related to the fact that the algorithm assigning the popularity value to each of the artists is not taking into account something that is underlying in the network, so it gives an hint on a possible way to improve Spotify recommendation algorithm.
By looking at the community found in the two Spotify cases (full and reduced networks) there are no big differences apart from the fact that from 18 they go down to 17 and they become less populated for the reduced one. The most interesting result is the AllMusic one where now there are 19 communities, against the 3 that could be identified just by looking at the graph. This could be a reflection of what one can deduce by looking at the clustering coefficient. By comparing the Spotify result with the AllMusic one, in the second case the fact that the communities were characterized by artists in the same genres (or subgenres) is lost. Once again, there is a mixing of Italian-singing and non-Italian-singin artists also in the communities, not only in the components.
Due to the small number of artists considered, I was not able to recreate the behaviour for the degree distributions as found in [3], so an exponential decay for human-generated (AllMusic) and power law for algorithm-generated (Spotify) networks, so a plan for the future could be to increase the number of nodes considered in order to get more significant results. In fact, what I've just considered were the first neighbours of the 100 initial nodes: one could go deeper in the analysis including higher order of neighbours and see how the results change including more nodes and links.

References

[1] https://newsroom.spotify.com/company-info/, Last accessed on 27/07/2023.

[2] Mark Newman. *Networks: An Introduction*. Oxford University Press, 2010. DOI: 10.1093/acprof:oso/9780199206650.001.0001.

[3] Pedro Cano et al. "Topology of music recommendation networks". In: *Chaos: An Interdisciplinary Journal of Nonlinear Science* 16.1 (2006). DOI: 10.1063/1.2137622. URL: https://doi.org/10.1063%2F1.2137622.

[4] Réka Albert and Albert-László Barabási. "Statistical mechanics of complex networks". In: *Reviews of Modern Physics* 74.1 (2002), pp. 47–97. DOI: 10.1103/revmodphys.74.47.