LastofNetworks: Studies of Artists Network on Last.fm

Biancamaria Bombino
b.bombino@studenti.unipi.it
Student ID: 561745

Guido Trentacapilli
g.trentacapilli@studenti.unipi.it
Student ID: 668551

Niccolo Seghieri
n.seghieri@studenti.unipi.it
Student ID: 666410

Pierfrancesco Benincasa
p.benincasa@studenti.unipi.it
Student ID: 657945

ABSTRACT

Last.fm is a music service that tracks the songs users listen to on different platforms and creates a detailed record of each user's musical activity. This allows users to see statistics on their listening habits, such as most listened artists and tracks, listening times, and much more. In this project, the main objective is to create a network based on the artists present and study the evolution and traits of this network incorporating different techniques of Social Network Analysis.[1]

KEYWORDS

Social Network Analysis, Last.fm, Music Interactions, Music Trends, Artists Interactions, Artists Collaborations, Genres Trends.

ACM Reference Format:
Biancamaria Bombino, Guido Trentacapilli, Niccolo Seghieri, and Pierfrancesco Benincasa. 2024. LastofNetworks: Studies of Artists Network on Last.fm. In . ACM, Pisa, PI, Italy, 11 pages.

[1]**Project Repositories**
Data Collection:
https://github.com/sna-unipi/sna-final-project-2024-lastofnetwork/tree/main/data_collection
Network Analysis and Manipulation:
https://github.com/sna-unipi/sna-final-project-2024-lastofnetwork/tree/main/network_analysis
Open Problem:
https://github.com/sna-unipi/sna-final-project-2024-lastofnetwork/tree/main/open_problem
Report:
https://github.com/sna-unipi/sna-final-project-2024-lastofnetwork/tree/main/report

SNA '24, 2023/24, University of Pisa, Italy

1 INTRODUCTION

Founded in 2002, **Last.fm**[2] uses a tracking system called *scrobbling* to record the tracks that users listen to across various platforms, such as Spotify, Apple Music, and local music players. Here are some key aspects of Last.fm:

- **Scrobbling**: This is the core feature of Last.fm. It tracks the songs users listen to on different platforms and creates a detailed log of users' musical activity. This allows people to see detailed statistics about their listening habits.
- **Music Recommendations**: Last.fm provides personalized recommendations for new artists, albums, and tracks that a user might like. These recommendations are based on users' musical tastes and those of other users with similar preferences.
- **Statistics and Analysis**: Last.fm offers a range of interesting statistics about music listening. People can view charts and lists of their most listened tracks, albums, and artists, compare their musical taste with their friends, and discover global music trends.
- **Community and Interaction**: Last.fm has a social component that allows users to connect, share their favorite music, and see what their friends are listening to. It is possible to follow other users, join music groups, and comment on artist profiles.
- **Events and Concerts**: Last.fm provides information about concerts and musical events, personalized based on the user's musical preferences and geographic location.
- **Integration with Other Platforms**: Last.fm can be integrated with other streaming platforms and music players through official plugins and apps, allowing for seamless scrobbling.

Metadata are crucial for the functioning of Last.fm. Each scrobbled track includes metadata such as the song title, artist name, album, release year, and musical genre. These data are used to catalog the music and provide personalized recommendations. Users can tag songs with keywords that describe the genre, mood, or other attributes of the music.

[2]https://www.last.fm/

These tags help to improve music recommendations and create personalized radio stations based on the user's musical interests. The purpose of this study is to conduct an analysis of tags that reflect an artist's characteristics, such as their musical genre, in order to obtain information about a possible community of performers. This information could be useful for studying potential collaborations among them, representing the artists and their connections based on tags with a graph. The nodes of the graph represent the musicians, and there is a connection between them if they share at least three out of the five considered tags. The edges are weighted based on the number of shared tags. During this study, all the characteristics of this network are explored, comparing it with some standard models: **Barabási-Albert** and **Erdős-Rényi**. In particular, a characteristic of the network is that it was not intentionally built for the purpose of creating connections, meaning it is not a network focused on social relationships. It is, in fact, based on common characteristics rather than direct social interactions, so it is more accurately described as a network of similarity or affinity.

Contribution

This paper as the follow contributions.

- Methodology for network construction (e.g., algorithms, data analysis techniques, etc.).
- Analysis of network characteristics (centrality, clustering, density, etc.) and comparison with standard network models.
- Identification of artist communities using community detection algorithms.
- Link Prediction for the formation of links.
- High-order Network Analysis.
- Potential deductions on collaborations among artists.
- Extensions with the open question and conclusions.

Paper Organisation

The report is organized as follows: Section 2 describes the procedure for downloading the data and constructing the network, while Section 3 presents a preliminary analysis of the network characteristics and a comparison with standard models. Sections 4, 5 and 6 describe the three chosen approaches, namely community detection, link prediction and higher-order network analysis. Finally, Section 8 discusses the open question selected by the team: "How are artists' communities structured in the tag-based network, what are the dynamics of centrality and interaction between musical genres, and what strategic roles do artists play within this network?".

2 DATA COLLECTION

The network considered was built with data extracted from Last.fm using the **Last.fm API**[3]. To obtain the data, this solution was preferred to the implementation of a web scraper for the site, as the platform allows you to register and obtain an **API key** which guarantees a high share of daily requests. In an initial phase, artists were downloaded to take their names, via *chart.getTopArtist*. Each request made via this method was set to return approximately 500 artists per request. In total the number of artists obtained is 10512. Subsequently, the *artist.getInfo* method was used, setting only the artist's name as a parameter. This method was carried out on all the artists obtained previously, and in this way the information of each person was obtained. In particular, the saved information of interest to us are tags. The final dataset *artisti_df.csv* contains artists' names and their five tags. In this phase, performers with less than five tags have been removed. The Table 1 summarizes the statistics of the data obtained.

Number of Artists	10512
Total Number of Tags	52560
Tags per Artist	5

Table 1: Dataset

An **Artist-to-Artist** network was built with the extracted data using artist as nodes. For the insertion of edges, reference was made to tags. An arc between two nodes was created if they shared at least three tags out of five. Also, the arc was weighing according to the number of shared tags, that is: if artists share three tags then the assigned weight is 1, if they share four tags the weight is 2 and if they share five tags the weight is equal to 3. In summary, the network scheme is as follows:

- **Artist**: Name of each artist present on Last.fm.
- **Tags**: Set of tags (max five) related to each artist. These tags can represent the genre, the country or nationality of the artist, the presence or not of the instruments, the type of exhibition, the vocalist gender and the profession (singer and/or songwriter).

On the analyzed platform, tags are labels associated with musical tracks, artists, or albums. These tags can be added by users, but they can also be generated automatically by algorithms or collected from external sources. To better understand future analyses, efforts were made to identify the most popular tags. In the Figure 1, it is possible to visualize that the most frequently mentioned tags, with a frequency exceeding 1000 occurrences, are: *rap, pop, indie, rock*. Meanwhile, other highly specific tags are characterized by much

[3]https://www.last.fm/api

lower frequencies, often even equal to 1 (such as otacore, neotrance). These results suggest a highly skewed distribution, with relatively few frequent tags dominating the dataset. The resulting network is a bidirectional network with 10042

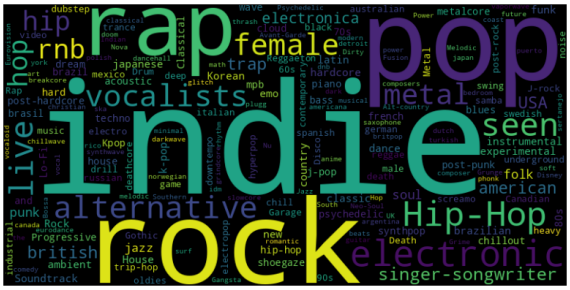

Figure 1: WordCloud of Tags

nodes and 725330 loops (of which 66 are self loops and are removed mainly in the link prediction task). These self loops likely represent noise in the data and in our specific case do not actually reflect a significant relationship between the artists; at most, they may indicate some consistency or cohesion in the artistic production of a particular artist, but nonetheless they are not relevant for our analyses. In Table 2 are shown the main features of the network of artists.

Number of nodes N	10042
Number of edges L	725264
LMAX	50415861
Average Degree <k>	144.446
Average Clustering Coefficient	0.514
Density d(G)	0.01438

Table 2: Artist-to-Artist Network

From this table you can see that the obtained graph is **scattered** due to:

- L « LMAX [4]
- <k> « N-1 [5]
- d(G) « 1 [6]

In fact, if the graph had been complete it would have had L = LMAX, average degree <k> = N - 1 and density of network d(G) = 1. In addition, the network falls under the **connected** regime, as the average grade is greater than **ln(N) = 9.21**. Further analysis will be presented in the following sections.

[4] LMAX = N(N+1)/2
[5] <k> = 2L/N
[6] d(G) = L/LMAX

3 NETWORK CHARACTERIZATION

The built network was compared with two other synthetic models whose parameters have been set to have the same number of nodes of the original network and a quantity similar to edges; in order to give a deeper perspective of how the network is composed, analyzing similarities and differences with the standard. The considered models, whose values arcs are shown in Table 3, are the **Barabási-Albert** with **m = 7** and the **Erdős-Rényi** in the version *Subcritical, Critical, Supercritical and Connected.*

BA	69951
ER Subcritical	4390
ER Critical	5038
ER Supercritical	9801
ER Connected	51032

Table 3: Nets Statistics

Degree Distribution Analysis

It's possible to observe the degree distribution of Last.fm network in the Figure 2. Observing this distribution, it is possible to notice a decreasing trend. The presence of nodes with very high degrees indicates the presence of hubs, which is crucial for the network's connectivity and robustness. The heterogeneous distribution of degrees could indicate that there are significant differences in the number of connections between various nodes. For the BA network shown in Figure 3, however, it can be stated that it perfectly follows a *Power-law*. This distribution decreases linearly, and there are fewer nodes with high degrees compared to the real network.

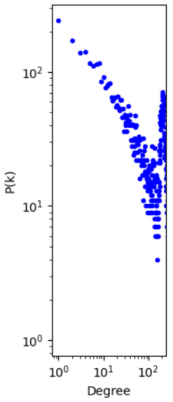

Figure 2: Last.fm Network

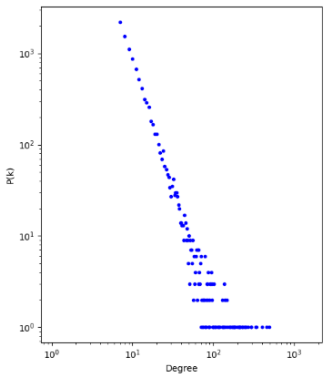

Figure 3: BA model

The network looks very different from the ER model, in fact for all its variants, the latter shows a distribution in

which there are no nodes with medium-high degrees. In the ER model, the degree distribution is close to a *Poisson* distribution for large networks. The absence of nodes with medium to high degrees could simply be an intrinsic feature of this distribution; since most nodes tend to have degrees close to average with a few nodes deviating significantly from this mean.

Connected Component Analysis

The network analysed consists of 30 connected components, of which the largest is 9960 nodes (99% of the total network) and for this reason it is a **giant component**. From this it is possible to deduce that there is a high interconnectedness between artists, and the presence of a giant component could facilitate the transmission of innovations and styles between them. In addition, this also suggests increased network resilience. The aritists not present in the larger component are probably less influential or those who deviate more from trends. As for the standard models, for the *BA* and the *connected ER* there was only one connected component; instead for the other types of ER there were many. In particular for *subcritical ER* have been found 5610 connected components of which the largest has a size equal to 92 nodes, and this is the direct result of the low probability of connection between nodes being p lower than the critical threshold. This low probability prevents the formation of large connected components, leading to a scattered and fragmented network. Instead, for *critical ER* there are 4964 connected components and the maximum has a size equal to 168 nodes, and for *supercritical ER* there are 1739 connected components and the maximum is formed by 7819 knots. This last results is obvious because p exceeds the critical threshold, and the ER network forms a giant component and the number of connected components decreases drastically.

Path Analysis

For the largest component of each network have been calculated the average shortest path and the diameter, shown in Table 4. It is evident from this outcome that the actual network has a diameter of 13. This suggests that there are nodes that require up to 13 steps (arcs) to be reached from each other. This value may indicate that the component is not extremely dense, otherwise there would be shorter paths between all nodes. It could be inferred that clusters or subcommunities have strong internal connections but are less connected to each other.
Assuming a possible collaborative aspect of the artists within this network, this result makes us understand the presence of a certain complexity and segmentation in the network of collaborations. The spread of influences, musical styles or work opportunities across the network could take up to 13 steps in the most extreme cases. This can have implications

	Diameter	Avg Shortest Path
Artists Network	13	3.55
BA model	5	3.36
ER Subcritical	23	9.77
ER Critical	35	24.12
ER Supercritical	30	12.18
ER Connected	7	4.23

Table 4: Path Statistics

on how quickly new trends or information are propagated among artists. Moreover the values more similar to the real network are assumed by the BA model and the connected ER, both composed by a single component. On the other hand, for the other synthetic nets, much higher diameter values were obtained and so they are less efficient. Finally, through the average shortest path value, the deductions made with the diameter can be confirmed. In fact, the BA and the connected ER have values similar to the average shortest path of the network of artists, while the remaining types of ER have values much higher that again indicate a greater dispersion of these networks.

Cluster Coefficient and Density Analysis

The global clustering coefficient of the graph is higher than that obtained in synthetic networks, as observable in Table 5. A clustering coefficient of 0.514 for the real network means that, on average, there is a 51.4% probability that the nearby nodes of a node are also connected to each other. This value indicates a relatively high level of clustering in the network, suggesting that nodes tend to form cohesive groups or communities. This may suggest that artists tend to collaborate in closed groups, that is: if a first artist collaborates with a second and the latter collaborates with a third, it is likely that the first and the third collaborate.

Real Network	0.514
BA	0.009
ER Subcritical	0.0
ER Critical	0.0015
ER Supercritical	0.00018
ER Connected	0.001

Table 5: Clustering Coefficients

Comparison with standard models suggests that the real network has a more complex, more resistant to random failures and organized structure than those generated by these synthetic models. The density of synthetic networks is lower than the density of the network of Last.fm. For the real network in fact the density is equal to 0.014, and to this value

is approached only the network of the standard model BA with a value equal to 0.0139. As for the types of ER, however, they have a lower density. The very low value of density in general for each network is justified by the presence of a number of arcs much less than the maximum possible number (*LMAX*) close to 50 million. In conclusion, the high clustering coefficient (compared with random network with same dimensions) and the low average shortest path (similar to random network with same dimensions) are a sign of a **small-world** network.

Centrality Analysis

The network has been analysed using methods based on different definitions of centrality: **degree based**, **connectivity based** (Eigenvector, pagerank) and **geometric based** (closeness, betweenness). This analysis was carried out on the first 10 nodes and the results are shown below in the Figure 4.

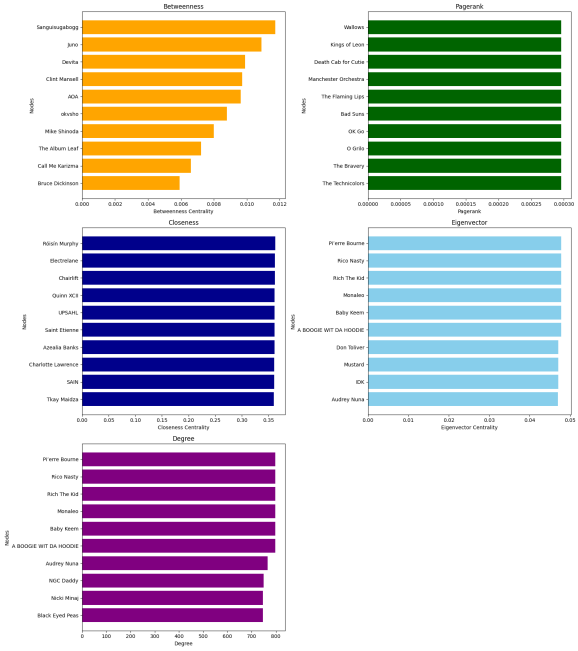

Figure 4: First 10 Artists for Centrality

Looking at the graphs of the first ten nodes obtained for different measures of centrality, it is possibile to make different deductions and comparisons. The knot with the highest centrality **betweenness** is *Sanguisugabogg*, followed by *Juno*, *Devita*, and so on. This metric indicates that these nodes play a crucial role in the connection of different parts of the graph. Although *Sanguisugabogg* has a very high betweenness centrality, indicating a crucial role as a bridge between different parts of the graph, it does not appear in graphs of other centralities. This leads to the conclusion that its role is specific to the connection between different communities rather than the number of direct links or global influence. The highest **pagerank** knots are *Wallows*, *Kings of Leon*, *Death Cab for Cutie*, etc. For this measure all the first ten nodes seem to have the same pagerank, and this high value suggests that these nodes are well connected and influential within the network. *Róisín Murphy*, *Electrelane*, *Chairlift* are the nodes with the highest **closeness** centrality, that is those that can spread information more quickly and efficiently. Nodes like *Pi'erre Bourne*, *Rico NASTY*, *Rich The Kid* have the highest values of **eigenvector** centrality. A high eigenvector centrality suggests that these nodes are not only well connected, but also connected to equally influential nodes. These three nodes also have the highest **degree** and for that reason they have many direct connections and are potentially important communication points in the network. In

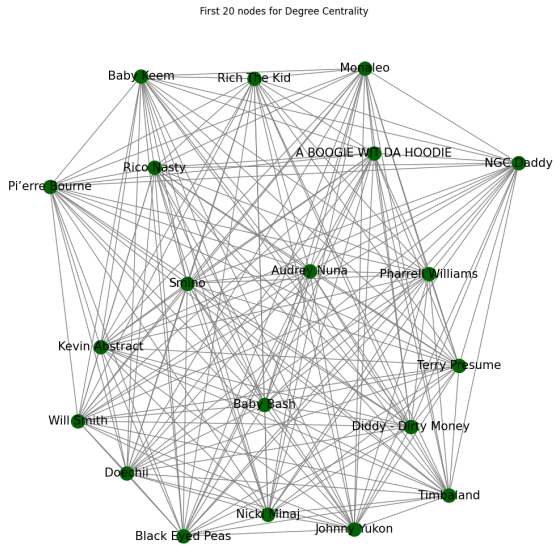

Figure 5: First 20 Artists for Degree Centrality

the Figure 5 it is possible to visualize the graph of the first 20 artists for degree centrality. For the first artists were also displayed the corresponding tags in order to understand the musical characteristics related to the most influential artists. It turns out that the tags most associated with them are *rap*, *hip-hop*, *trap* and *seen live*. From this it is possible to conclude that evidently these musical genres are particularly inclined to collaborations and featuring. Moreover, this may reflect a cultural dominance of these genres in the current music scene, and may be the main trendsetters and innovators in the music scene. Lastly, artists with this style could be seen as key partners. Finally, the assortative mixing was analysed. In particular, the value of the **assortativity coefficient** is

0.685. This value close to 1 indicates that artists tend to connect with others who have a degree similar to theirs. For example, nodes with many connections tend to connect with others with many connections, and nodes with few connections tend to connect with others with few connections. As a result it is likely that groups of nodes with high degrees form dense communities, while nodes with low degrees form more sparse communities.

4 TASK 1: COMMUNITY DETECTION

In this Section the goal is to identify the communities of artists present in the network with different algorithms and compare the results. The techniques used are **Label Propagation, Louvain, Leiden** and **Infomap**. For the **Louvain** algorithm a grid search has been made to look for the optimal parameters, and the obtained configuration is as follows: *'resolution': 0.9, 'randomize': None.*
With this randomize parameter setting, the Louvain algorithm always starts from the same initial configuration, allowing a more accurate comparison with other methods. The Table 6 shows the results of the evaluation metrics obtained by each technique.

	AID	IED	COND	MOD	N_Comm
Label Prop.	9.77	0.72	0.28	0.45	163
Louvain	24.48	0.69	0.059	0.61	45
Leiden	27.24	0.69	0.061	0.61	44
Infomap	12.98	0.64	0.39	0.60	204

Table 6: Results Community Discovery Algorithms

From the table, it is apparent that *Infomap* is the algorithm that has detected a greater number of communities, presumably due to its ability to detect very detailed communities, which is an approach based on information theory. *Leiden* and *Louvain* instead tend to produce fewer communities than other algorithms because they are optimized to maximize the modularity of network division in coherent and meaningful communities. In fact, the latter report higher values of modularity and consequently this suggests that the communities found are more consistent and well separated than the others. To these higher values of modularity are also associated two values of conductance lower than *Infomap* and *Label Propagation*. Lower values of conductance indicate that few connections cross the community boundary, meaning that communities are well isolated from the rest of the network. As far as the IED values are concerned, there is no clear difference between the various techniques, so probably the communities identified by each algorithm have a similar density of internal links, and consequently the same cohesion within the communities. Instead, *Leiden* reports the maximum value of AID followed by *Louvain*, which again confirms that their communities have nodes that are more strongly connected to each other. So it can be said that **Louvain** is the method that reported the best results in identifying communities. For this reason, the composition of the first two larger communities, formed by 2891 and 1349 artists respectively, were visualized through the wordcloud in the Figure 6.

Figure 6: Louvain's first two largest communities

For the evaluation of the results, **coverge** was also compared. The highest value was reported by *Label Propagation*, with a node coverage of 93%. This suggests that the communities found by the label propagation better reflect the structure of the network, with strongly linked nodes included within the same communities. While the lowest coverage value was obtained by *Infomap* which covered only 69% of the nodes. This may indicate that the communities found are not representative of the actual structure of the network, with groups of strongly connected nodes that have not been included in coherent communities. In this way, the Infomap coverge further explains the high number of communities achieved. In this regard, the Figures 7 and 8 depict the cardinality of communities generated by Label Propagation and Infomap.

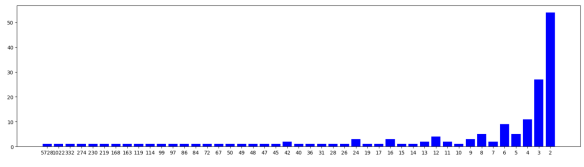

Figure 7: Cardinality of communities in Label Propagation

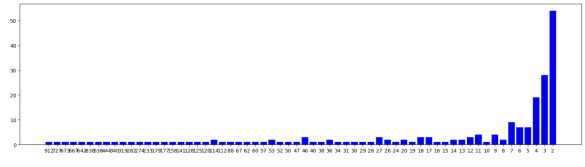

Figure 8: Cardinality of communities in Infomap

The only thing in common between the four methods is that all identify **crisp**-type communities, where each node belongs to at most one community.

5 TASK 2: LINK PREDICTION

The second task of this project is the Link Prediction [1], useful to predict the existence or future formation of links between nodes. Different metrics from different families were used for the experiments conducted, namely:

- **Neighborhood-based**: whose measurements assign a score between two nodes x and y, determined by the neighbors who have in common the two nodes. *CommonNeighbours*, *AdamicAdar* and *Jaccard* have been tested in this category.
- **Ranking**: considering the *SimRank* measure, which assigns score based on similarity of neighbors

Three metrics were calculated for each technique: *accuracy*, *precision* and *recall*. The Table 7 contains the outcomes obtained and it is possible to notice the difference between these methods.

	Precision	Recall	Accuracy
CommonNeighbours	0.2098	0.2098	0.9953
AdamicAdar	0.2264	0.2264	0.9954
Jaccard	0.2774	0.2774	0.9957
SimRank	0.0537	0.0537	0.9944

Table 7: Link Prediction Methods outcomes

Observing the values it is evident that *Jaccard* is the metric that reports the best values of precision and recall, indicating that it is the most effective in predicting the correct links between artists. *SimRank*'s precision and recall are significantly lower, suggesting that this method is less effective in predicting correct links in this network. In a large network, like the one with 10042 nodes (artists), the total number of possible connections is very large. However, in many real networks, only a small fraction of these possible connections actually exist. The high accuracy obtained in all cases could be misleading due to the high proportion of non-collections in the network, and this value indicates that most model predictions are correct. Nonetheless, in a scattered network, most pairs of nodes are actually not connected. Thus, a model that predicts mainly non-linkages could still achieve high accuracy. Finally, to predict future collaborations, the link prediction algorithm based on the distance of nodes in the graph was used. In this algorithm, existing arcs in the prediction were excluded, focusing only on potential new links. For the results the distance scores were calculated for all possible new links and the five links with the worse (greater) distance scores were selected, indicating the less likely links:

- **Kienan Auton - Hendrika Whellams**: 1.5
- **Viola Lafferty - Abel McClymont**: 1.2
- **Vally Chidwick - Kort Cuffe**: 1.2
- **Thane Greet - Niall McVeighty**: 1.2
- **Tabbitha Gilogly - Idette Wyllcock**: 1.2

So according to the algorithm, *Kienan Auton* and *Hendrika Whellams* have a distance of 1.5, making them potentially less likely to form a link than the other results.

6 TASK 3: HIGH-ORDER NETWORK ANALYSIS

As a third task it was decided to carry out the Higher-order Network Analysis, belonging to the Network Manipulation. The hypergraph represents an extension of the artist graph where hyperarcs can connect more than two nodes, providing a more flexible model for representing complex relationships. The Figures 9 and 10 depict a partial representation of the hypergraph and its dual, but for issues of space and readability only the first five artists were represented.

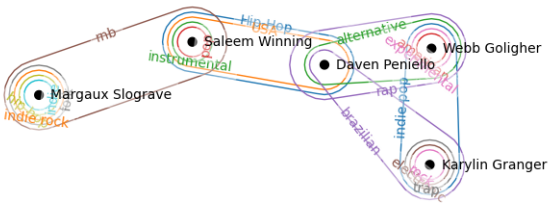

Figure 9: Hypergraph of the first five artists

For example observing such figures, *Saleem Winning* and *Daven Peniello* seem to be central with many connections to other artists through various tags. This indicates that these artists could share many common interests or influences with other artists in the network. *Margaux Slograve* and *Karylin Granger* are linked to fewer other artists, suggesting that their collaborations or influences are more specialized or less widespread.

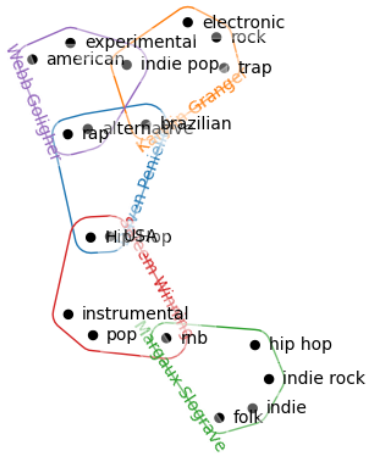

Figure 10: Dual Hypergraph of the first five artists

In the dual hypergraph context, tags become nodes and artists become the hyperarcs that connect these nodes. Tags that are linked to many other tags through various artists indicate music styles or influences that are widely shared among artists in the network. From the image, it is visible that tags like *hip hop* and *alternative* are central, connecting many artists. These tags probably represent key musical influences or styles prevalent among artists. Later, to better understand the structure, the hypergraph with 1000 nodes is shown in the Figure 11.

Figure 11: Hypergraph of 1000 artists

From this representation makes it evident the presence of a part completely separated from the other indicating that there are two connected components disjoint in the graph. This means that the group of nodes at the top has no connection to the larger group at the bottom. This could be due to the fact that the artists of the right group share tags that are not present among the artists of the left group, creating a clear division. In addition, the left part of the graph is very dense and compact, suggesting that most tags are highly interconnected through various artists. This indicates a music community with a wide range of shared influences. In order to better understand the structure of the network and the hypotheses deduced from the previous hypergraphs, further analyses were carried out. First of all the **density** value for hypergraphs was found, considering all nodes, which is equal to 0.00178. Such a low density in the hypergraph indicates that most artists share a relatively small number of tags with other performers. This suggests that the network is not very cohesive; artists are connected in a sparse way, with few groups sharing many tags. Tags tend to be quite specific and probably do not appear frequently in common combinations. The network of artists is probably fragmented into many small components. To get an idea of the fragmentation of the network, the **s-components** and their dimensions have been analyzed to understand how large the communities are and how scattered they are. In the Figures 12 and 13 the distribution of s-components for the complete hypergraphs is shown.

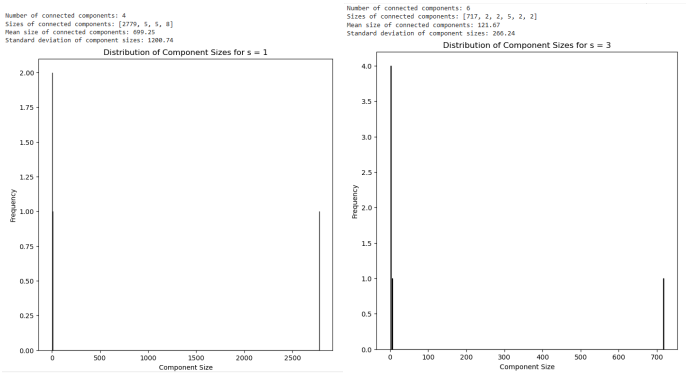

Figure 12: s-components distribution for s=1 and s=3

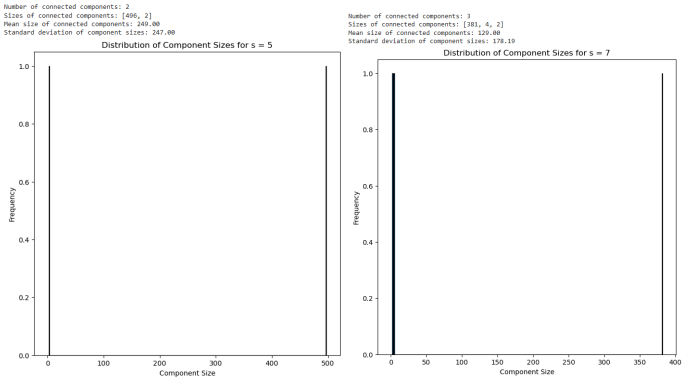

Figure 13: s-components distribution for s=5 and s=7

From these images it is possible to deduce that for **s = 1** the network has only 4 connected components, with a very large dominant component (2779 nodes). For **s = 3** the number

of components increases to 6, suggesting that requiring at least three connections between nodes further fragments the network. For **s = 5** the number drops to 2, instead for **s = 7** rearranged to 3. With the increase of *s*, the size of the largest component decreases. This is predictable because requiring multiple common connections, only a subset of nodes can remain connected. The network is more cohesive with *s = 1*, showing a large main component connecting most nodes. Cohesion decreases as *s* increase, showing increasing fragmentation. Finally, the presence of a large component with *s = 1* suggests that many artists are only connected through a few common tags. For memory errors, the s-components for the full dual were not possible to calculate. For this reason a sample of 1000 knots was used in this case. From the results obtained it is observed that at low thresholds **(s = 1)**, the tags are highly connected because you get a component of size equal to 999, suggesting that the music tags tend to share many artists in common and creating a very dense network. For **s = 3**, there is a large core of strongly connected tags (component of 791 nodes) and many smaller components. This suggests the existence of a central group of popular or widely used tags, with various niche tags forming separate groups. By increasing *s* to 5, the network splits into tag pairs, indicating that only very specific and closely related tags remain connected. With **s = 7**, there are no connections, showing that there are no groups of tags so strongly related that they share seven or more artists. Finally, the **degree centrality** of the complete dual hypergraph was checked, so as to see the number of artists that are associated with each tag. In particular the top five tags with a higher associated artists value are:
[('pop', 2120), ('electronic', 1950), ('seen live', 1870), ('indie', 1860), ('rock', 1767)].
So these five tags are among the most common in the network.

7 OPEN QUESTION

In this Section several questions were asked to better understand the structure of the network. Initially we tried to understand how are distributed the musical genres within the communities found through the best algorithm (Louvain) obtained in the Community Detection Section. To conduct this analysis were selected the first four largest communities, and the results obtained are represented in the Figure 14.
From the image it is noted that the **first community** seems to have a strong presence of *British* and *electronic* genres and the presence of the *seen live* tag suggests that many artists are popular for their live performances. The **second community** is dominated by *indie* and *rock* genres. Again, *seen live* is prominent, indicating a strong element of live performance. The *Lo-Fi* genre also highlights a preference for more intimate and acoustic styles. The **third community** is

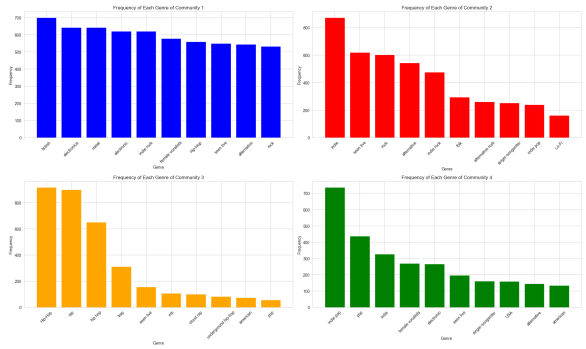

Figure 14: Genres in the first four largest communities of Louvain

clearly oriented towards *hip-hop* and *rap*, with various subgenres such as *trap* and *cloud rap*. The presence of seen live is less noticeable and for that reason it could be centered on artists who are more popular online or through recordings. The **fourth community** is strongly characterized by *indie pop* and *pop*, with a significant presence of indie artists and female vocalists. The variety of tags like *USA* and *American* indicates a strong geographical component. Also here, *seen live* is present, indicating a good popularity for live performances. The *degree centrality* and *betweeness* were analysed below for each community, taking into account the first top ten nodes resulting from these metrics. These calculations have been performed to understand if there are specific tags that tend to form clusters of artists and which tags are the most common among larger clusters. For the first community the resulting tags were *british* and *electronic*, for the second *indie* and *seen live*, for the third *rap* and for the fourth *pop*. Subsequently, efforts were made to identify the performers most likely to act as **bridges** between different communities. To do this the number of external arcs for each node of each community was calculated and the first ten musicians with the greatest value were displayed. For the first community the person with more external connections is *Azealia Banks*, for the second is *Fleurie*, for the third is *Example* and for the fourth is *Janna Creggan*. From this it is possible to deduce that these four artists probably facilitate the flow of information and collaborations between different groups and they could have a greater visibility and influence in the overall network because they are not isolated in their community but interact with multiple groups. Finally, these people could belong to more genres or musical styles, showing an artistic versatility that allows them to connect with a wider and varied audience. In fact, for example, some tags associated with *Azealia Banks* are *electronic*, *rap* and *hip-hop*. The latter are the dominant genres in three different communities. Continuing with the theme of musical categories, an additional question asked concerns which musical genres tend to have

more central artists in the network, and whether there is a correlation between the musical type and the measures of centrality (i.e. degree, betweenness and closeness centrality).

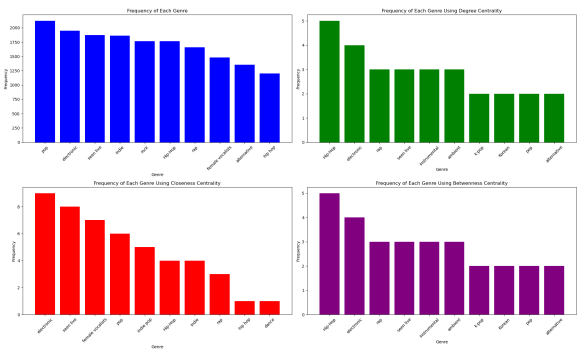

Figure 15: Genres frequency for each centrality metrics

From the graphs in the Figure 15, it emerges that **Hip-Hop** and **electronic** are particularly prominent in terms of **Degree Centrality**, which indicates that artists of these genres tend to have many direct connections in the network. These are specific tags that are more effective in creating links. **Electronic** also has a very high frequency in terms of **Closeness Centrality**, indicating that artists of this genre are close to all other nodes, facilitating rapid access and communication across the network. In addition, the presence of **female vocalists** and **seen live** in this category communicates that these genres have artists who are strategically well positioned to influence many other performers in the network. The music categories with the highest **Betweenness** are **Hip-Hop** and **electronic**, signaling that the musicians belonging to this field act as crucial points for communication between different parts of the network. On the other hand, genres like *k-pop*, *Korean*, *pop* and *alternative* have a lesser presence in terms of *Degree* and *Betweenness Centrality*. As a result, the artists associated with them are less directly connected and less involved in connecting different parts of the network. Following with the experiments, has been tried to understand if the artists with a greater diversity of tags (i.e. that belong to more styles) tend to be more central and occupy more strategic positions in the network. For this purpose a dictionary of macro genres was created and the number of artists belonging to one or more macro categories was calculated. As a result, most performers (8769) belong to at least five macro genres and there is no artist specializing in a single type of music. In addition, it has been noted that nodes belonging to at least five styles are efficient for the dissemination of information and could also be strategic (if removed they can fragment the network), being characterized by a high closeness and betweenness. Instead, no artist belonging to more than four genres is among the most active nodes. In fact the nodes with the highest degree centrality are those belonging to maximum four categories, leading to infer that the musicians who act as hubs are however multi-genres. To further investigate any strategic performers, the central nodes were removed randomly. From this experiment an artist named **Juno** was highlighted, as it was noticed that with the removal of this knot from the network the number of components increased from 30 to 31. The representation of this situation is depicted in the Figure 16. From the subgraph the strategic position of *Juno* is highlighted, this being the only possible connection with **lor2mg**. This is probably due to the fact that this artist only shares with *Juno* the minimum threshold of required tags, set during the creation of the network, for the creation of the links. In fact *lor2mg* contains tags too specific or particular such as *cute* [7] and *screamo* [8].

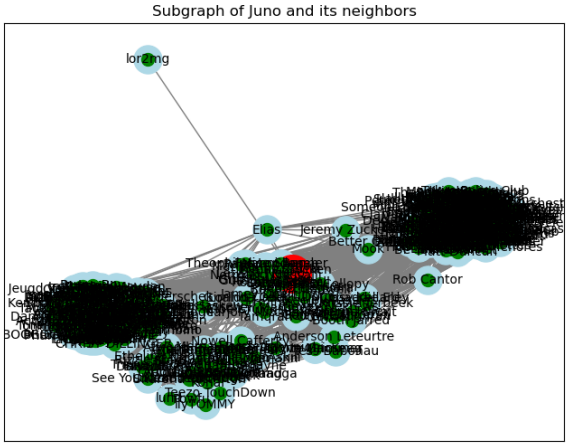

Figure 16: Rappresentation of "Juno" matter

8 CONCLUSIONS

In conclusion, the network built allowed us to successfully obtain information regarding possible communities and collaboration between artists on Last.fm. In particular, community detection algorithms revealed distinct groups of performers who share more connections among themselves than with those outside their community. These communities often correspond to similar genres or collaboration patterns. In addition, the network turns out to be quite robust and resilient, as seen with the removal of some central nodes. Finally, a possible future work can be the temporal analysis, investigating how the artist network evolves over time in order to provide insights into the dynamics of musical trends and collaborations.

[7] related to the style or atmosphere of a song
[8] subgenre of punk evolved mainly from hardcore punk in the early nineties

REFERENCES

[1] Jon M. Kleinberg David Liben-Nowell. 2003. he link prediction problem for social networks. (2003).