Universiteit Leiden
The Netherlands

Opleiding Informatica

Uncovering Influential Artists and Musical Communities in an Artist Co-Follower Network

Zoë Abhelakh

Supervisors:
Dr. Akrati Saxena & Dr. Frank Takes

BACHELOR THESIS

Leiden Institute of Advanced Computer Science (LIACS)
www.liacs.leidenuniv.nl

01/07/2023

Abstract

Through the rise of the internet people express interest by following their favourite artists on social media. We create a network of artists that are linked through having common followers. This *artists' co-follower network*, constructed by combining the Discogs database and Twitter, allows for social network analysis of the artists and their relations. We investigate influential artists in the network by employing centrality measures to determine an artist's role in the network and the correlation with the musical attributes of the artists. We assess the existence and strength of traditional music communities often used to define artists by genre or style in the network. In addition, we investigate whether these communities align with densely found partitions by utilizing a community detection algorithm. The research contributes insights into audience similarity and looking beyond traditionally defined music communities.

Contents

1 Introduction

The rise of social media platforms has drastically changed the way people express their interests and engage with content. This transformation has also greatly impacted the music industry, enabling artists to promote their music and connect with their audiences in new ways. Unlike traditional promotion methods, such as through radio, live shows and interviews, social media allows artists to interact directly with their fans. The simple act of following or liking an artist's page has become enough to demonstrate one's support for an artist. A key metric for being popular has shifted from selling the most albums to being a key figure on social media. In addition to creating this new dimension for people to interact with others and post content on, social media platforms also record every interaction, generating valuable information on a large scale for researchers to explore [11].

Drawing inspiration from the theory of *homophily* by Mcpherson et al. [18], which assert that: "contact between similar people occurs at a higher rate than among dissimilar people". We can draw the notion that people who are similar to each other in music preferences are also more likely to follow the same group of artists than they would with someone with a very dissimilar music preference. This concept leads us to explore the *artists' co-follower network*, a social network of artists, where the number of common followers between artists determines the relationship between them. This network can then be used to explore the inter-connectedness of artists on Twitter, where the similarity of their followers influences relationships.

By leveraging social network analysis techniques, we aim to explore the artists' co-follower network created from Twitter data. This allows us to identify important artists in the networks through influence and means of connecting groups of artists. Additionally, we investigate communities of artists and explore if traditional methods of categorising artists by genre or style play a role in the manifestation of these communities.

1.1 Terminology

This section provides a brief overview of key terms used throughout this thesis.

1.1.1 Social networks

Social networks are complex systems of individuals, groups, or organisations connected through various relationships or interactions. Social network analysis is the structural approach in social science that focuses on studying these complex systems and the relationships therein. Using graph theory, social structures can be represented in a general structural model [7]. The social actors in this general structural model can be characterised as nodes, while edges represent their interactions or relationships with one another.

A social network is represented as a graph $G = (V, E)$, where V is the set of nodes in the network. The connections between these nodes are captured by the set of interactions, denoted by E and often referred to as edges or links. An edge is denoted as $\{u, w\}$ for $u, w \in V$. A network is directed or undirected and weighted or unweighted. In a directed network, edges indicate a one-way relationship between nodes, while in an undirected network, edges indicate a two-way relationship

between nodes. In an unweighted network, all edges have equal importance, while in a weighted network, edges are assigned a value or weight that reflects the distance or strength between two nodes. Figure 1 shows an undirected, weighted network with five nodes and five edges.

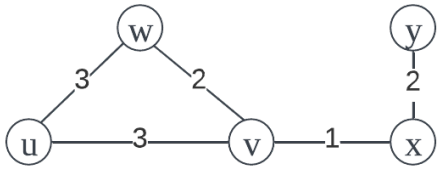


Figure 1: Undirected, weighted network

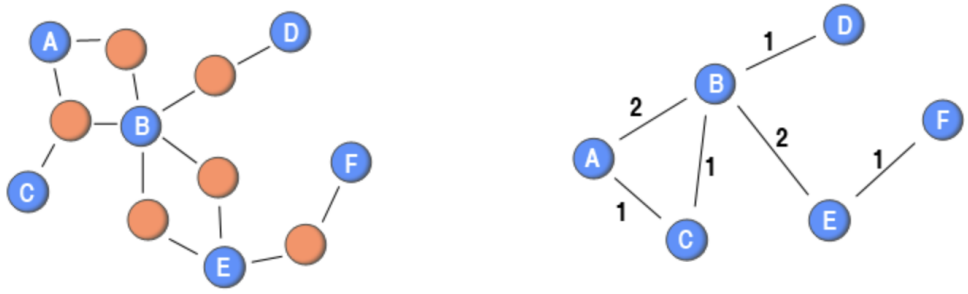
Other terms used in this thesis:
Giant component: The largest group of connected nodes in a network.
Density: The proportion of possible edges in the network that are actually present.
Triangles: A group of three nodes that are all interconnected to one another.

1.1.2 Two-mode networks

Section 1.1.1 introduces the generic structure of a social network. A two-mode network is a type of social network that consists of two distinct types of nodes, each connected exclusively to nodes of the other type [21]. For example, in the context of attending classes, a two-mode network can be formed with two types of nodes: students and classes. Each student attends multiple classes, forming connections between student nodes and class nodes. However, a class cannot attend another class, and a student cannot attend another student, resulting in connections solely forming between students and classes. Figure 2a illustrates two-mode network with links between the blue and orange nodes, but not within each set.

1.1.3 Network projection

The transformation of a two-mode network into a one-mode network involves creating a projection of one set of nodes, denoted as set X, and creating links between nodes of X based on if they share a common node from set Y [21]. To not lose all valuable information in the projection process, the number of common shared nodes functions as the edge weight w_{uv} where $u, v \in X$ in the projected graph. This is not only done to directly visualise the interactions between the chosen node set more intuitively [28]. Network projections are made to simplify the network structure and focus on the interactions between a specific set of nodes. Figure 2 shows the transformation of a two-mode network into a one-mode network. It illustrates that both node a and node b have connections to the same two orange nodes, indicating a weight of 2 for the link between them due to their shared nodes in set Y (orange nodes). Figure 2b depicts the projection on the blue nodes, where the number of common orange nodes determines the edge weight.



(a) Two-mode network (b) One-mode network

Figure 2: Projection of two-mode network to one-mode network
source: `https://toreopsahl.com` [21]

1.2 Research question

The following research question will be investigated in this thesis.

What insights can we gain about artists by analyzing their co-follower network on Twitter?

We can break down this research question into sub-questions:

1. How can we derive an artists' co-follower network from Twitter data?
2. How does network analysis reveal the most influential artists and their role in bridging different groups of artists?
3. What communities can be detected in the network, and what do they signal about music, genres, artists and their followers?

1.3 Thesis overview

The current section provides an introduction of the thesis; Section 2 lays out a review of related work to this research. Section 3 describes the data collection method, which is then used to apply the selected methods to in Section 4. Section 5 details the application of the methods. The implications, limitations, and potential future research are discussed in Section 6.

2 Related work

In this section, we describe prior work in the field of social network analysis related to the topics presented in this thesis.

2.1 Ethnomusicology in social network analysis

Social network analysis has been a useful tool due to its properties to analyse large-scale datasets in ethnomusicology, i.e., the study of music in social networks. The extensive musical landscape, made up of music, artists, record labels and more, has been studied through social network analysis. Collaboration networks in the music industry by Budner and Grahl [**?**] created insights into artists co-appearing on songs/albums and who have the highest connectivity in this industry. McAndrew et al. [16] explored the social network of British composers and showed that one's social network influences the work of a composer, as they could serve as direct or indirect inspiration for one's work. They also linked success to the centrality of a node [24]. Wi et al. [27] examined the influence of the social relationship between DJs by choosing songs to play in their setlists and found that artists only choose music from their exclusive social circle.

Lambiotte and Ausloos [13] analysed correlations between music groups of different genres and described the complexity of attributing genres to music groups. The analysis also showed that music listeners categorise the same music group in various genres and styles.
A study by Vlegels and Lievens [26] analysed a two-mode responder-artist network to examine the limitations of predetermined genre lists in measuring patterns in music taste and cultural omnivorousness. The study takes a critical stance on using genres as static values with rigid boundaries that do not consider the ever-changing properties of genres. The study revealed that artist clusters defy traditional genre boundaries. This questions the idea of categorising artists based on genre affiliations and highlights the importance of recognising different artist clusters within genres.

Considering the complexity of categorising artists with genres as described by Lambiotte and Ausloos [13], it is crucial to recognise how it affects the artists' co-follower network. In our dataset, a single genre from the Discogs database is attributed to an artist. The extent to which this impacts the network can not be estimated. However, the music preference of Twitter users diffuses the borders of genres as they follow artists they like and not the genres they like. Artists being grouped in the network is now more nuanced than just the genre of the artist. Additionally, this study will make use of predefined lists for genre, style, and label to see if there exists a relationship between how connected artists of the same genre are.

The findings from Vlegels and Lievens' study [26] on the limitations of predetermined genre lists provide valuable insights and can be taken into account when interpreting the result. By incorporating a community detection approach, this research will also aim to identify clusters of artists that are tightly connected, transcending traditional genre boundaries. This approach allows for a more nuanced understanding of music preferences, going beyond predefined categories and capturing the inherent complexity and interplay between artists by incorporating the ideas introduced by Vlegels and Lievens' work.

2.2 Identifying key players in a social network

Identifying key players or influential individuals has been a topic of interest in the field of social network analysis.

One approach, as described by Himelboim [11], focuses on centrality measures [24] to identify influential users on social media. Hubs, characterised by their high degree of centrality, are social actors with many connections in the network. They are central to information flow and attention within social media networks. On the other hand, bridges are actors that connect different parts of the network that would otherwise not be connected. They fill structural holes in the network and enable the flow of information between disconnected groups or nodes. Bridges have the power to control the spread of information, making them essential for reaching new audiences.
On the other hand, Borgatti [5] argues that these methods are not optimal for solving the key player problem. He proposes two methods to identify key players, known as `KPP-NEG` and `KPP-POS`. Key Player Problem/Positive (`KPP-POS`) metrics aim to identify optimally positioned actors that diffuse information maximally. This approach focuses on the connectivity of a key player and how they are embedded within the network and is most closely related to degree centrality. Key Player Problem/Negative (`KPP-NEG`) metrics aim to identify actors whose removal would break up the network maximally and is most closely related to betweenness centrality.

In the context of our co-follower network of artists, the identification of hubs is often straightforward, as it can be identified by the follower count of an artist. However, identifying bridges between different communities within the network becomes crucial for information diffusion and collaboration. Partnering with a bridging artist provides access to a diverse audience that is not necessarily your own, helping to increase an artist's visibility and expand their reach. Recognising and collaborating with bridge nodes in the network is a strategic way for artists to grow their influence in the music industry.

3 Data

This section outlines the process of data extraction from different sources, data refinement, and data integration to establish a solid foundation for conducting analyses. In this research, a total of three data sources were used to create a social network of artists on Twitter.

The artist data used in this research to create a social network was sourced from the Discogs database, a crowd-sourced discography database providing extensive information on music releases and additional metadata [10]. We accessed the publicly available data dump from January 1, 2023, which encompasses information on over 8 million artists and 15 million releases [1]. It is important to note that the definition of 'artists' in this dataset extends beyond just the lead singers; it also includes composers, audio technicians, and other individuals involved in the music production process. To ensure data reliability and compatibility with other sources, we addressed inconsistencies and missing values in the data through a process of data cleaning.

The Twitter network data used in this study was collected by Kwak et al. in 2010 [12] and obtained for our research from the Stanford Network Analysis Project (SNAP) [15]. The data collection processes involved two methods: The first method used a breadth-first search approach, starting at Perez Hilton's profile and systematically crawling the Twitter users she followed and those who followed her. The second method involved gathering user profiles who had tweeted about a popular topic within a specific time frame. This Twitter snapshot, using both collection methods, includes 41.7 million Twitter user IDs and 1.47 billion directed relations between Twitter users.

To integrate the Discogs database with the Twitter snapshot, a connection needed to be established between the artist names in Discogs and their corresponding Twitter IDs. This linking process was done using the Twitter API [2]. Figure 3 illustrates a visual representation of this integration method, where the Twitter name listed in the Discogs database was cross-referenced using the Twitter API. If a match was found, the artist's Twitter ID was obtained and linked to the Twitter snapshot data.

In total, 90,000 artists (1.1%) out of 8 million artists in the Discogs dataset were found to be present on Twitter. 14,467 artists of the Discogs dataset were present in the Twitter snapshot in 2010. Figure 3 also illustrates the incorporation of two sets of domain-specific attributes. The first domain includes attributes sourced from the Discogs database, including genre, style, and label, which are directly related to the artist's musical work. The second domain contains Twitter attributes, such as follower count and verified status, which were retrieved during the data collection process that took place between March 9 and March 22 of 2023. Therefore, the second-domain attribute values reflect information of the Twitter profile within that time frame.

[1] https://discogs-data-dumps.s3.us-west-2.amazonaws.com/index.html?prefix=data/2023/
[2] https://developer.twitter.com/en/docs/twitter-api

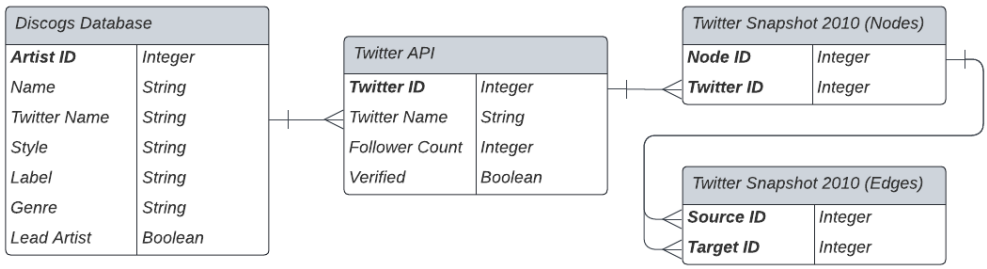


Figure 3: Data integration diagram: connecting Discogs database, Twitter API, and Twitter snapshot

4 Methodology

In this Section, we describe the setup of the research and which software was used, then methods discussed that were applied to answer the research questions described in Section 1.2

4.1 Research setup

The data obtained in Section 3 was stored in a PostgreSQL [3] database. Additionally, to extract the required data, data extraction methods were defined in SQL. The network analysis methods were applied in Python using the Networkx [2] package. Visualisations of the network were made using Gephi [4], while statistical graphics were generated using the Matplotlib [1] and Seaborn [3] libraries. The implementation of the pruning filter described in Section 4.3 was carried out using the author's Git repository [5].

4.2 Network formation

Section 1.1.2 introduces the definition of a social network. In Section 1.1.2 and Section 1.1.3 explains the concept of two-mode networks and how to perform network projection on it. For this research, a social network of artists was constructed, named an artist's co-follower network, based on the network projection of a two-mode network.

Section 3 described the data collection and processing methods. The resulting dataset is then split into a node list, which represents the artists with additional metadata, and an edge list of around 1.4 billion edges. While the node list only holds artists, the edge list contains the following pattern of all Twitter users in the Twitter snapshot. This information is needed to calculate the number of shared followers between artists. Consequently, the edge list holds two types of nodes: artist nodes and user nodes. In this case, the artist nodes should only be connected to the non-artist nodes and vice versa.

[3] https://www.postgresql.org
[4] https://gephi.org
[5] https://github.com/naviddianati/GraphPruning

We use the following logic to determine for every user if they are an artist or non-artist: An artist is always a Twitter user. However, a Twitter user is not always an artist. Let V be the set of nodes in the edge list with every Twitter user $v \in V$. Let U (users) and A (artists) be two subsets of V. A Twitter user $\in A$ if the Twitter user is present in the node list of artists. All Twitter users not present in the node list are automatically part subset U. This results in two distinct types of nodes in the edge list: U (users) and A (artists).
To ensure the network in this research is a true two-mode network, we enforced the following rules to the edges:

- A directed edge can only exist between a node $u \in U$ and a node $a \in A$. In other words, for any edge (u, a), we have $u \in U$ and $a \in A$.
- For any Twitter user $u \in U$, there exist at least two artists $a1, a2, \in A$ such that u follows both $a1$ and $a2$.

The enforcement of these rules do not only create a true two-mode network but also eliminate relationships among Twitter users that are unimportant to our research, resulting in a drastic reduction of edges by 99.88%, which causes a significant saving in computational time and resources.

We applied network projection to our network on the set of artists, where the number of shared Twitter users is denoted as the edge weight. This explains the second rule in Section 4.2, as a Twitter user who follows at most one artist will not participate in any overlap with another artist.

Let $G = (A, U, E)$ be a two-mode network where A is the set of artists, U is the set of Twitter users, and E is the set of edges between A and U. Network projection is applied to obtain a one-mode network $G' = (A', E')$ where A' is the set of artists in G and E' is the set of edges between artists in A' based on the number of shared Twitter users. The edge weight between two artists $a_i, a_j \in A'$ is defined as the number of Twitter users who follow both artists, denoted as $w_{i,j}$. The second rule in Section 4.2 is based on the fact that if a Twitter user follows at most one artist, then there is no edge to be created between that artist and another artist, resulting in $w_{i,j} = 0$ for all $a_i, a_j \in A'$.

The resulting one-mode network, thus called the artists' co-follower network, contains 14,467 nodes and 9,718,426 edges. The network is weighted and undirected. This increase in edges can be attributed to the projection creating edges between nodes that did not have a direct edge in the original two-mode network but were indirectly connected through a shared node in the other set of nodes.

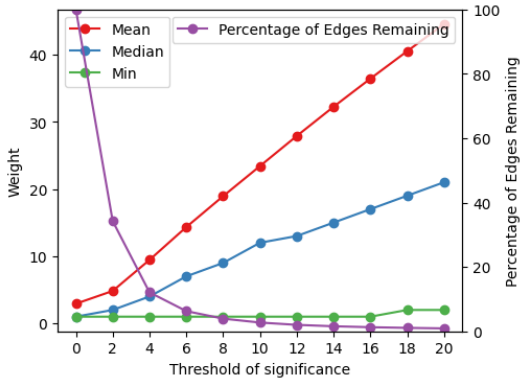
4.3 Network thresholding

The artists' co-follower network obtained in this study contains numerous edges, making it challenging to interpret and computationally intensive. Pruning algorithms aim to reduce the number of obfuscating edges to reveal an unknown, underlying structure [6] and reduce computational costs. Various methods have been proposed to address this issue, such as weight-based thresholding, where edges below a certain weight are removed, and proportional thresholding, which retains a specific percentage of edges [8]. However, weight- and proportional thresholding can potentially disadvantage smaller artists in the co-follower network. This issue arises due to their relatively

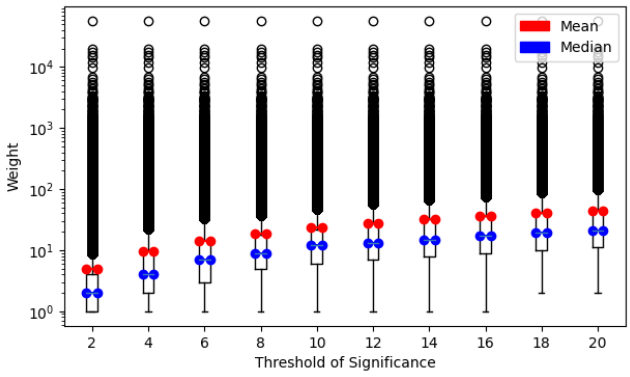
smaller following, which inherently decreases their likelihood of overlapping audiences with other artists. Consequently, these artists' edges may be removed and they become isolated from the network.
To account for this disadvantage, Dianati [6] introduces the marginal likelihood filter (MLF), which assesses the statistical significance of an observed edge weight in a graph by comparing them to a null model. The null model preserves the graph's degree sequence and weight sequence and assigns edges randomly based on node degrees. The null model allows the computation of the probability mass function for edge weights. By comparing the observed weights to the null model, a p-value can be calculated to determine their statistical significance. The MLF filter has been shown to extract a larger and sparser giant component compared to thresholding methods based solely on absolute edge weights.

The statistical measures of the network at various significance thresholds applied by the MLF filter are illustrated in Figure 4. The plot in Figure 4a reveals a significant decrease in the percentage of remaining edges, indicating a substantial proportion of edges that are deemed of low significance to the overall network structure. However, the mean and median display minimal changes, as evidenced by Figure 4b. This observation suggests that, with increasing thresholds, lower weight values are progressively eliminated. It is worth noting that this elimination process occurs at a slower pace compared to previously mentioned weight-based thresholding methods.

The threshold of 16 was selected as it allowed for the removal of a significant number of edges before reaching the point where the edges with the lowest weight would be eliminated, as indicated by the green line in Figure 4a. Furthermore, the mean and median values show only slight increases in Figure 4b. Consequently, this threshold resulted in the removal of 98.72% of edges. These results suggest that the selected threshold effectively pruned the network by removing less significant edges while preserving the structure of the remaining network.



(a) Mean, median, min, and percentage of edges vs. threshold



(b) Weight distribution with mean, median, and threshold

Figure 4: A comparison of statistical measures when applying different thresholds of significance

4.4 Attribute aggregation

The initial Discogs dataset contains information at two levels: the artist level and the music releases level. The relationship between the two can be described as a many-to-many relationship, where a single artist can produce multiple releases, and multiple artists can contribute to a release. To reduce the dimensionality of the dataset and facilitate connections between attributes at the music release level and the artist level, we employed attribute aggregation. This process involved aggregating attributes based on their most frequent occurrences for each artist. Genre, style and label were retrieved using the method described in Appendix A to aggregate information from the music releases and applying it to the corresponding artist.However, it is essential to recognize that reducing a many-to-many relationship to a one-to-one relationship inevitably results in some information loss. This study acknowledges that an artist's musical attributes may evolve over time, and thus a single aggregated value does not fully capture the nuances of their artistic identity.

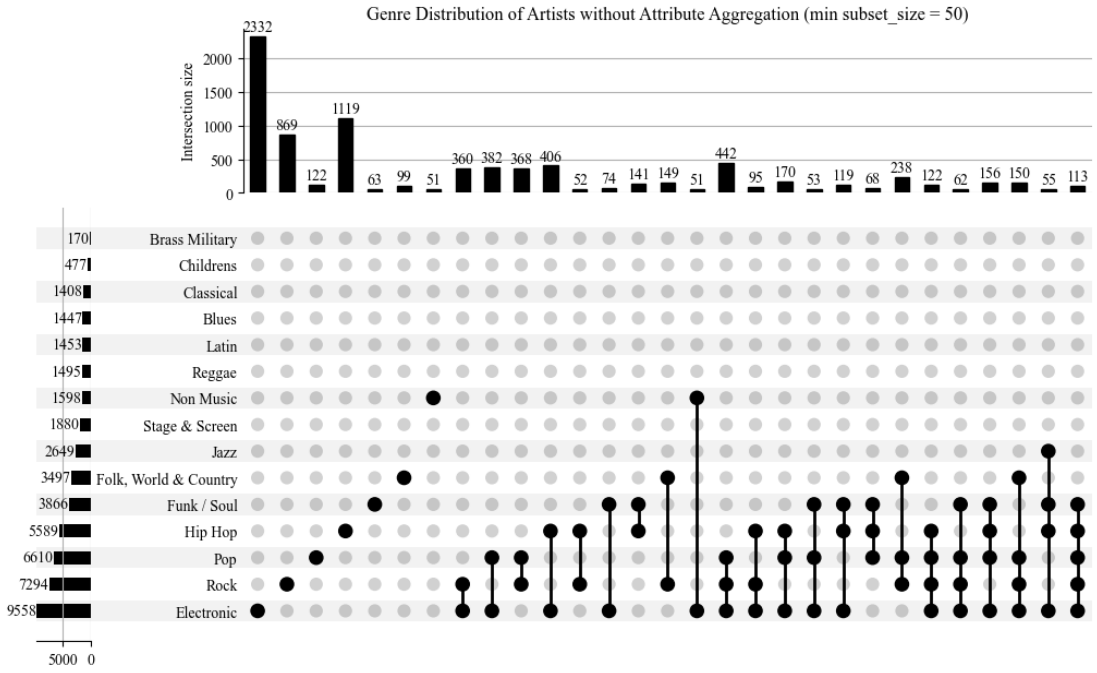


Figure 5: Genre distribution of artists in artists' co-follower network without attribute aggregation

During the initial analysis, it was observed that the approach of aggregating attributes based on their most frequent occurrences did not adequately capture the different combinations of genres in which artists often released music. This limitation is illustrated in Figure 5, which presents histograms of intersection sizes for genre sets containing at least 50 artists. These sets are generated based on whether artists have ever released music within those genres. The figure demonstrates that while the three largest genre sets are accurately represented as single genres, the subsequent larger sets consist of combinations of different genres. This indicates that with genre aggregation of only the most occurring genre of an artist, we would lose valuable information. Furthermore, separating artists into single genres may inaccurately depict artists' communities.

To address this limitation, we introduce a refined aggregation method for genres. The method allows for more granular analysis and aims to understand if introducing more subcategories in genre creates a more accurate and reliable result. The refined aggregation method involves selecting all genre sets that exceed 1% of all nodes in the network, corresponding to a threshold of 144 nodes per set. Sets with more than one genre are transformed into new sets that combine the names of the genres. In total, six new genres will be created in addition to the fifteen existing ones. As a result, the size of genres Electronic, Pop, Rock and Folk, World & Country will drop. The new genres introduced are as follows:

- Electronic, Pop, Rock
- Electronic, Hip Hop
- Electronic, Pop
- Pop, Rock
- Electronic, Rock
- Folk, World & Country, Pop, Rock

Lastly, relying solely on a predetermined list of genres may not adequately capture the complexities of an artist's musical preferences. This limitation has been recognised in previous studies, including Vlegels and Lievens [26]. A more comprehensive overview of their findings is presented in the Section 2.1.

4.5 Influential artists

To gain insights into the structure of social networks, it is important to explore the existence of influential entities who hold a disproportionate amount of influence due to their position in the network. This position can be characterised by a high degree of connectivity or the ability to connect otherwise disconnected groups in the network [11]. Social network analysis allows the quantification of entities that have significant importance within a network [5], which, in the context of our research, can be attributed to an artist's popularity. In this section, two methods of finding influential artists are discussed based on the idea of centrality and bridge positions in the network.

Multiple methods will be applied to assess the centrality of a node in a network and investigate potential correlations with domain-specific or node-specific attributes. It is important to consider that the edges have weights, which impacts the calculation of some centrality measures. The strength of the relationships, indicating the extent of overlapping audiences, must be considered in these calculations.

4.5.1 Degree centrality

Degree centrality is a measure based on the number of connections a node has. Nodes with a high degree of centrality serve as hubs in the network and have many connections [11]. The degree centrality is unweighted and only captures the number of links, not the weight they might hold. In this research, the degree of centrality indicates how many different artists an artist shares followers.

The degree centrality is calculated by taking the number of links a node has and normalising that value by dividing it by $(n-1)$, where n is the total number of nodes in the network.

4.5.2 Closeness centrality

Closeness centrality measures how close a node is to all other nodes in terms of distance. This is done by calculating the average shortest path length from a given node to all other nodes in the network. A higher closeness centrality indicates a more central node [9].

In the context of weighted networks, algorithms designed to find the shortest path take the weight between nodes as the distance. Therefore, in networks where the weight represents the strength of the relationship between nodes, the inverse of the weight must be considered to identify the shortest path for a given node [22].

4.5.3 Betweenness centrality

Betweenness centrality calculates the number of times a given node was present in the shortest path between two other nodes in the network [9]. Betweenness centrality, like closeness centrality, uses the calculation of the shortest path between two nodes, so the weight must be inversed to account for the strength of the path as described in Section 4.5.2.

Nodes with relative high betweenness centrality play an essential role in connecting distinct parts of the network that lack direct relationships [11]. In the context of artists' co-follower networks, betweenness centrality assesses an artist's ability to act as a bridge between different types of artists who have no overlapping followers. It reflects the artist's capacity to facilitate communication and control the flow of information between these diverse artists' communities.

4.5.4 Eigenvector centrality

Eigenvector centrality measures a node's importance in the network based on its connections to other important nodes. The measure is based on the idea that being close to an important node makes oneself more important [11]. For eigenvector centrality, the weight is already interpreted as strength between two nodes. In the context of the artists' co-follower network, this means that that artist with high eigenvector centrality have an overlapping audience with artists that have high centrality measures.

4.5.5 Finding key players

Section 2.2 introduces the idea of key players as entities that act as hubs or bridges in a network. The concept of key players, discussed in Section 2.2, in a network involves identifying entities that act as hubs or bridges. Hubs are nodes with many connections, while bridges connect different parts of the network that would otherwise be disconnected. Removing bridge nodes can significantly impact network connectivity.

The method proposed by Borgatti is noted to be closely related to betweenness centrality. To identify key players and bridges in our artists' co-follower network, we calculate the betweenness centrality for each node as described in Section 4.5.3. Nodes with high betweenness centrality are potential bridge nodes that play a crucial role in connecting distinct parts of the network. Upon removal of these nodes, they could potentially cause the network to become fragmentated or overall

less well connected.

To assess the impact of these key players on the network structure, we systematically remove nodes with the highest betweenness centrality. By observing the resulting changes in the number of network components and path lengths, we gain insights into the significance of these influential nodes and their contributions to the overall connectivity and information flow within the network.

4.6 Community structure

To gain a deeper understanding of the network structure, we can examine it at a higher level beyond individual nodes. Community structure refers to the partitioning of the network into groups of nodes, which are more densely interconnected with each other and less with the rest of the network.

In our study, we aim to investigate whether the community structure of the artists' co-follower network is influenced by the musical attributes assigned to artists. Specifically, we aim to explore if the community structure in the network aligns with the traditional categorization of the musical landscape, such as by genre and style. We examine whether artists with similar attributes, such as being labeled with the same genre or style, tend to be more strongly connected to each other within the network. This analysis allows us to assess the extent to which the network reflects established categorizations in the music industry.

4.6.1 Attribute assortativity

In the context of this research, we applied assortative mixing to explore the presence of attribute-based communities within the artists' co-follower network. Assortative mixing, as described by Newman [20], refers to the tendency for nodes in a network to be connected to other nodes that share similar characteristics. The attributes are all categorical. Thus, we can use discrete attribute mixing. To quantify the level of assortative mixing in an unweighted network. Newman [20] defined an assortativity coefficient that is calculated as:

$$r = \frac{\sum_i e_{ii} - \sum_i a_i b_i}{1 - \sum_i a_i b_i} \tag{1}$$

where

- e_{ii} is the fraction of edges that connect nodes of the same type i
- $a_i b_i$ is the product of fractions of each type of end of an edge that is connected to nodes of type i.
- The numerator of the equation measures the difference between the observed assortativity mixing and the expected mixing due to random connections.
- The denominator serves as a normalisation factor to the numerator by dividing by the inverse of the expected mixing due to random connections.
- r is the assortativity coefficient, which lies in the range $-1 \leq r \leq 1$.

Since the assortativity coefficient measure is unweighted, this treats all artists' relationships equally. However, using the concept of homophily, we expect artists with the same attribute values to be more similar to each other and have a stronger relationship. To investigate this, we introduce a weight threshold to remove low-weight links in the network. This allows us to re-evaluate the assortativity, aiming to determine if low weights obfuscate the results and stronger links would yield a higher attribute assortativity.

4.6.2 Affinity

Affinity refers to the measure of the correlation between a social network's structure and its entities' attributes. It captures the tendency of entities to connect and form links with other entities that share similar attributes. It can be used to measure if nodes with the same attribute are more likely to be connected to each other than they would be by chance. Affinity, as described by Mislove [19], quantifies this phenomenon as:

$$S_a = \frac{|(i,j) \in E : a_i = a_j|}{|E|} \tag{2}$$

where E represents the set of all links in the network and a_i the value of attribute a for an artist i. S_a represents the fraction of links for which artists share the same value for attribute a. This value is then divided by the fraction of links representing attribute a for a random graph with the same distribution of degree values. The random graph was generated 1000 times for an accurate distribution of values in a random graph. The outcome of this, affinity, ranges between 0 to ∞ and denotes the ratio of the fraction of links between attribute-sharing artists relative to what would be expected in a random graph. Thus, an affinity greater than 1 indicates that links are positively correlated with artist attributes. The affinity measure is unweighted and only accounts for the existence of a link and not the weight thereof.

4.7 Community detection

We utilise community detection algorithms that uncover communities that are not determined by the available attributes but instead reveal underlying patterns. The presence of music genre information for the nodes in the network allows us to explore whether the communities identified by the algorithms predominantly consist of individuals who release the same genre of music. The data consists of fifteen genres; an additional six are included by applying the attribute refinement method introduced in Section 4.4. Additionally, there are 298 styles, which serve as subtypes of the genres. The introduction of the refined genre attribute was specifically intended to investigate whether the community partitioning based on the genre would lead to a more nuanced distribution of genre values across multiple communities.

4.7.1 Louvain method

Louvain's method for community detection is a two-phase algorithm that aims to identify dense groups of nodes in a network. In the first phase, all nodes are assigned their own community. During the second phase, nodes are moved between communities to increase the modularity score. The algorithm systematically evaluates potential moves for each node and selects the one that maximises

the increase in modularity. This process is repeated until no further improvement in modularity can be achieved [4].

4.7.2 Leiden algorithm

The Leiden algorithm is an extension of the Louvain method by using refinement steps to improve the quality of the communities detected by ensuring that communities are well-connected. The algorithm aims to optimise a modified version of the modularity measure that considers the size of the communities [25]. A different phase is added to take this into account, called the refinement phase between both previous phases. This phase in the Leiden algorithm improves the quality of the community structure obtained by the Louvain algorithm. It allows for detecting smaller and more densely connected communities in the network.

4.7.3 Infomap algorithm

The InfoMap algorithm is a community detection algorithm that uses the concept of random walks to identify groups of nodes that are likely to be visited together, forming communities within the network. The algorithm's core objective is to minimize the map equation, a measure that quantifies the information required to encode the random walks on the network. By minimizing this equation, InfoMap seeks to find the optimal partitioning of the network into communities.

During the simulation, random walks are performed on the network, and nodes that frequently occur together in these walks are considered more likely to belong to the same community. By iteratively optimizing the map equation, the algorithm identifies the optimal representation of the network's community structure [23].

4.8 Robustness of network

To ensure the reliability and validity of our findings, we conducted a robustness check to assess the stability and consistency of our results. Specifically, we focused on examining the robustness of our network analysis to different methods of thresholding.

In our initial analysis, we applied in marginal likelihood filter in Section 4.3 to construct the network by retaining only the most significant edges. However, to evaluate the generalisability of our findings beyond this specific network construction approach, we applied a different thresholding method based on absolute edge weight.

In this alternative approach, we aimed to prune roughly the same number of edges (99%) by choosing a threshold that cuts out all links below a certain weight. The obtained network will be compared to the original network on various network statistics. This weight thresholding robustness analysis allows us to determine the extent to which our network analysis results are influenced by the specific choice of thresholding method.

5 Results

This section presents the results obtained in response to the research questions outlined in Section 1.2. First, we examine the network characteristics in Section 5.1. Subsequently, we apply the methods described in Section 4 to address the research questions and analyze the obtained results.

5.1 Network statistics

Table 1 presents the statistics of the artists' co-follower network on Twitter. The table shows that the network is characterized by a high level of fragmentation, consisting of 6165 components.However, a significant portion of the network's nodes (56.3%) and edges (99.9%) are concentrated in the giant component. This indicates that the remaining components in the network are isolated or have very few connections. This outcome aligns with expectations, considering the pruning filter applied in Section 4.3. The filter aimed to eliminate insignificant edges that could obfuscated the network structure. As a result, the nodes not connected to the giant component anymore were considered to have insignificant edges to the network structure. The rest of the analyses ran on the largest connected component of the network.

Network measure	**Values**
Number of nodes	14,467
Number of edges	124,218
Average degree	17.17
Number of components	6,165
Nodes in giant component	8,151
Edges in giant component	124,058
Average edge weight	36.45
Density	0.0037
Number of triangles	1,713,226
Average clustering coefficient	$9.97703 * 10^{-5}$

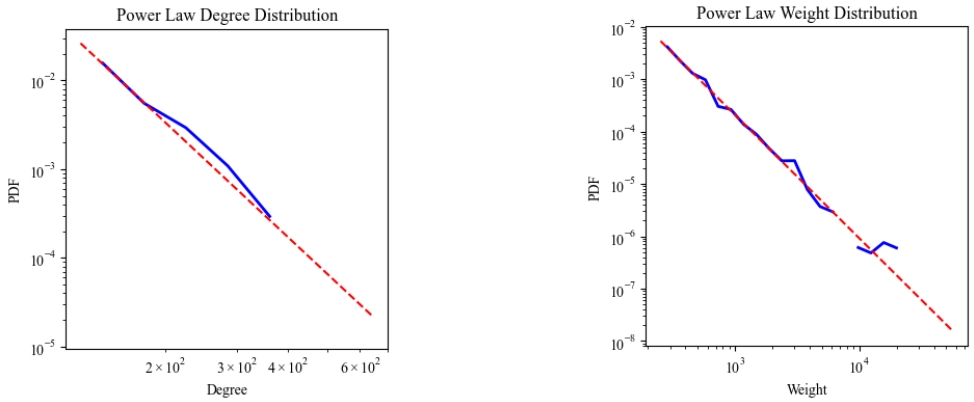
Table 1: Network statistics of artists' co-follower network

The average clustering coefficient is significantly low, which suggests that, on average, the nodes in the network are not highly connected to each other. This is in contrast with the high number of triangles. This indicates that a small group of nodes is highly connected, forming many triangles, while the rest of the network is not. This can be attributed to the Marginal Likelihood Filter implemented in Section 4.3, which aims to extract a sparser giant component. However, as illustrated by Dianati [6], this directly leads to a significant drop in the clustering coefficient, especially when removing 98.72% of edges.

5.1.1 Power laws

As discussed in Section 4.5 and derived from the network statistics in Table 1, a small group with a disproportionate amount of influence exists in the network. These networks do not follow a normal distribution but are heavily skewed towards this small group. This phenomenon is called the power law [11].

Another indication of the existence of a power law in our network, and an explanation for the low clustering coefficient but a high number of triangles, can be observed in Figure 6. This figure showcases the probability that a given node has a certain degree (Figure 6a) or weight (Figure 6b) displayed in a logarithmic graph. The power law distribution in Figure 6a illustrates that a small group of artists have a high degree, but most of the artists have low degrees. Also, Figure 6b shows that the weight distribution of the network follows a power law, with a small number of links having high weights, while most links between artists have low weights.



(a) Power law degree distribution

(b) Power law weight distribution

Figure 6: The power law distribution of degree and weight

5.2 Attribute statistics

Other than network-specific measures, the network also includes domain-specific attributes that may provide insights into relationships within the network. Figure 7 displays the distribution of values across the music-related attributes. Firstly, genre and refined genres, introduced in Section 4.4, showcase that the introduction of new combined genres has decreased the size of the most prominent genres, as expected. The distribution of style and label largely follow the same path of a skewed distribution. However, there is a vast difference in the number of values and the size of the values. The average number of nodes that the same style has is 27.3, while the average number of nodes with the same label is 1.5. The distributions indicate that all the attributes are not equally represented in the network.
Furthermore, Figure 8 shows the distribution of genres across the entire Discogs dataset, artists present in Twitter snapshot and artists' co-follower network used for analysis. This gives an in-

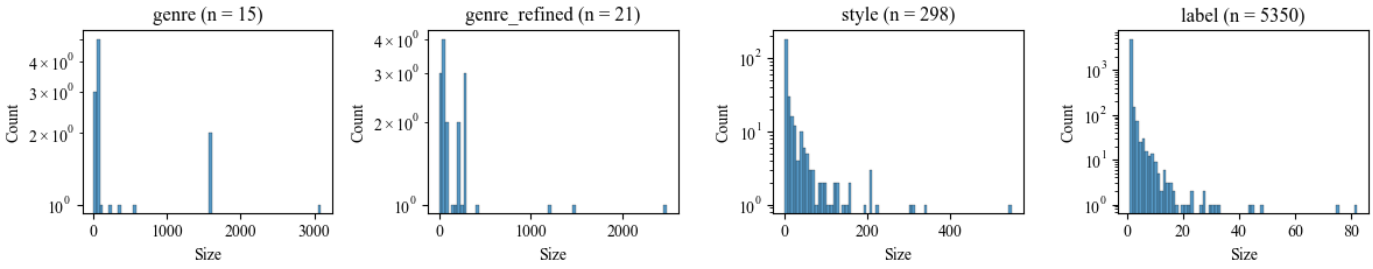


Figure 7: Distribution of attributes for genre, refined genres, style, and label

dication how well the genre distribution is accurately depicted in the artists' co-follower network compared to the data sources. The figure illustrates that the three largest genres by size make up over 60% of all Discogs database data. The genre distribution in the Discogs database and Twitter snapshot compare fairly; we notice that the two biggest genres are slightly overrepresented in the Twitter snapshot. The artists present in this research indicate a strong resemblance to the entire artist population.

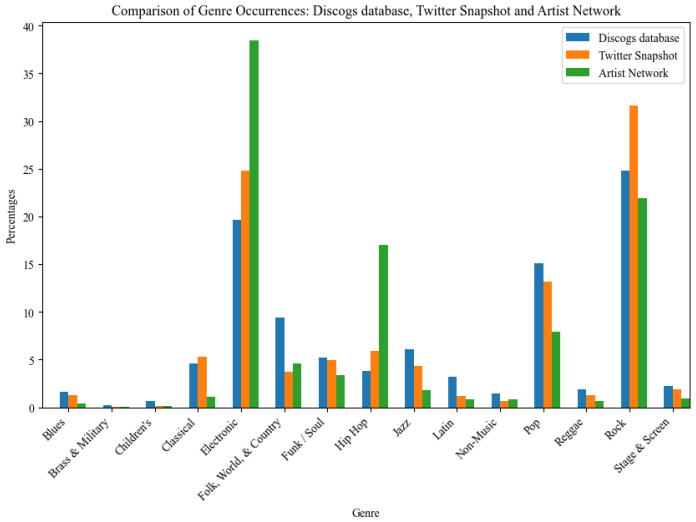


Figure 8: Comparison of genre occurences: Discogs database, Twitter snapshot of artists and artists' co-follower network

However, the comparison with the artists' co-follower network shows that 'Electronic' and 'Hip Hop' have respectively increased almost 2 and 3 times in percentage. This results from attribute aggregation (Section 4.4, which assigns each artist only one genre. Figure 9 shows the distribution of refined genres compared to the original genres. We see that the extremities of certain genres are slightly reduced and assigned to the newly introduced values. However, 'Electronic' and 'Hip Hop' remain largely overrepresented in the artists' co-follower network.

Figure 10a illustrates the fifteen largest styles by size. The colouring is based on the genre that they belong to. These 15 out of 298 styles cover 41.9% of all nodes in the giant component. The figure

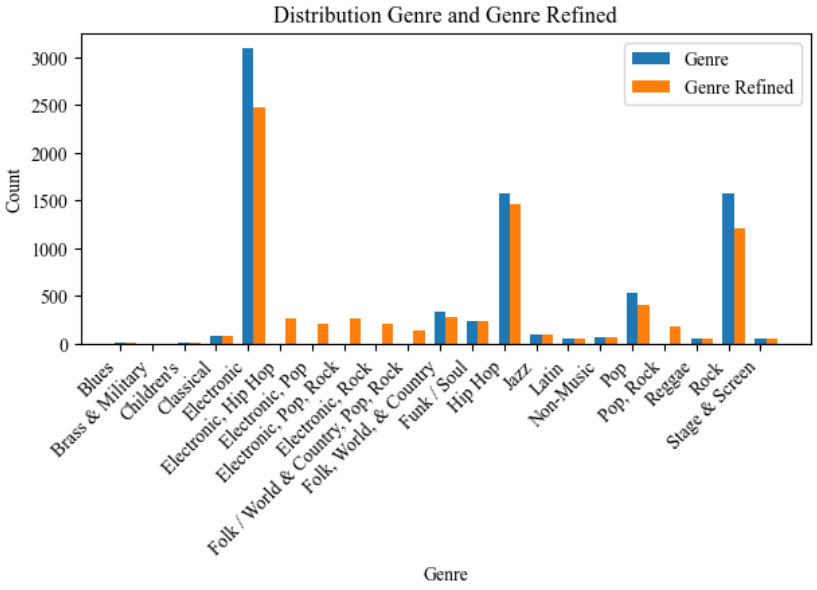
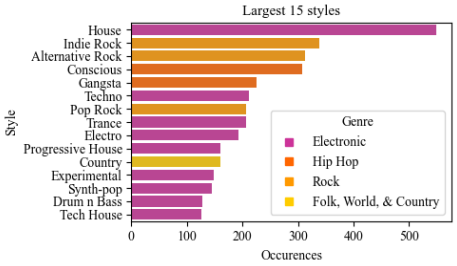


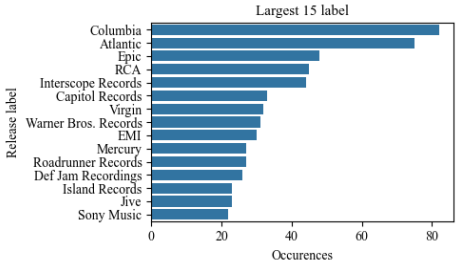
Figure 9: Distribution of genre and refined genres in the network

shows that the largest styles almost entirely belong to the largest genres in the network, which is explained considering style is below genre in the music classification hierarchy. However, style, like genre, suffers inherently from the attribute's dynamic nature. Styles can be a niche within a genre but also result from a crossover between genres. A good example is 'Pop Rock', which could be classified as 'Pop' or 'Rock'.

Figure 10b showcase the 15 most prominent release labels. Together they cover only 7.0% of nodes in the giant component. Thus the remaining 5335 labels cover the other 7589 nodes, resulting in an average assignment of 1.4 nodes per label.



(a) The 15 largest styles in the network



(b) The 15 largest release Labels in the network

Figure 10: Largest 15 values occuring in style and label attribute

5.3 Finding influential artists

Previous Section 5.1.1 reveals the existence of a small group with much influence. We can gain insights into who might be in these groups and their role in the network using centrality measures.

Additionally, we investigate whether a correlation exists between centrality and domain-specific attributes in the network.

Table 2 showcases the top five artists based on different centrality measures. It is important to note that the data used in this analysis dates back from 2010. Thus, the list of famous artists may have varied significantly since then. The first noticeable thing is that Barack Obama ranks highest on all centrality measurements indicating that he is a highly influential figure with an overall central position in the network. Due to his position as President of the United States at the time, his central position in the network and the overlap in followers cannot only be attributed to his artistic endeavours but is likely the result of his social status and political influence.

	Degree Centrality	**Closeness Centrality**	**Betweenness Centrality**	**Eigenvector Centrality**
1	Barack Obama	Barack Obama	Barack Obama	Barack Obama
2	Beat Butcha	Britney Spears	Lars Behrenroth	Britney Spears
3	Yoko Ono	Yoko Ono	MC hammer	Jimmy Eat World
4	Question	Jimmy Eat World	Yukmouth	Yoko Ono
5	Bjork	Imogen Heap	Taro Fumizono	Imogen Heap

Table 2: Top 5 artists of different network measures

The degree centrality in Table 2 reveals the list of artists with the highest number of links. However, this set of artists only reaches 1298 unique nodes, equating to 9.0% of all nodes, although their total reach extends to 2154. The overlap in the reached nodes suggests that there may be similarities or associations between the artists that lead them to largely follow the same group of artists. Furthermore, the average degree of the top five artists is 430.8, which is 2409% more than the average degree of nodes in the entire network. The average degree of the reached nodes (77.3) is also significantly higher, with a 350% increase compared to the network average. These findings suggest that the top-ranking artist and their first-reached nodes have a strong presence and connectivity within the network and represent the small group of artists with a high degree in the Power Law Degree. The significant difference in average degree further highlights the artists' prominence and potential hierarchical positions within the network structure.

The data also shows that the same set of artists are present for both closeness centrality and eigenvector centrality, indicating that the artists are located close to other artists in the network and have connections to other highly influential artists. Overall, being present in the top five of both measurements indicates a strong presence in the network. On the other hand, the set of artists ranked highest for betweenness centrality, not taking Barack Obama into account, is distinct and not present in the other centrality measures. The artists might not be closely located to many nodes or connected to influential nodes, but they form the bridge between different parts of the network, which without their role, would not be as close to each other.

5.3.1 Correlation of network measures

In order to gain a deeper understanding of the network topology, we explore the potential relationships between centrality measures of the artists. This analysis aims to determine if the observed

relationships highlighted in Table 2 are present throughout the entire network and not limited to the most influential artists.

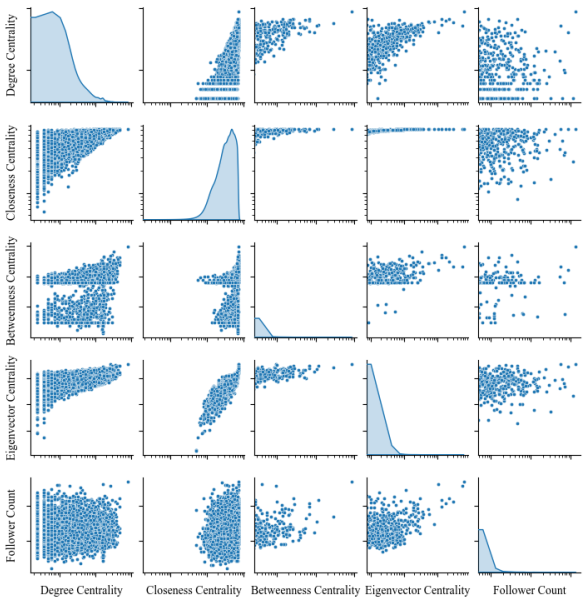


Figure 11: Scatterplot matrix of centrality measures

Figure 11 shows the correlation between various centrality measures, including the artist's follower count, which reflects their current degree. Most measures do not display any significant correlation. However, we observe a moderate positive correlation (0.60) between closeness centrality and degree centrality, which implies that an artist with many shared followers with other artists also tends to have a shorter path to other artists.

Additionally, eigenvector centrality and betweenness centrality display a moderate positive correlation (0.65). This suggests that artists who are connected to other higher influential artists are also present in the shortest path between two artists. This indicates that well-connected artists have direct influence and act as important intermediaries in facilitating communication or information flow. It is important to note that sharing a follower with another influential entity in this co-follower network does not necessarily indicate one's own influence. The relationships between artists in this network cannot be directly measured as influence.

Also, the follower count displays a slightly moderate positive correlation with eigenvector centrality (0.50) and betweenness centrality (0.47). This implies that artists that have a high follower count in 2023 were slightly more likely to be connected to other well-connected artists and were present in the shortest path between other artists. These measures, eigenvector centrality and betweenness centrality could then be used as potential indicators of future popularity. Interestingly, there is a minimal positive correlation between degree centrality and follower count, suggesting that a high degree centrality does not necessarily translate to a higher follower count at present. However, this observation may be influenced by the data collection method. The snowball method used might not have captured all the relationships Twitter users had at the time of data collection.

In conclusion, while all the correlations are positive, the relationships found in the top 5 artists do not seem to extend to the rest of the network completely. The rest of the results can be found in Table 8 of Appendix C

5.3.2 Correlation of domain-specific attributes

Investigating the relationship between network measures and domain-specific attributes provides insights into how an artist's position in the network is influenced by the specific attributes assigned to them. By examining the correlation between these attributes and network measures, we can gain a deeper understanding of the factors that contribute to an artist's placement and prominence within the network.

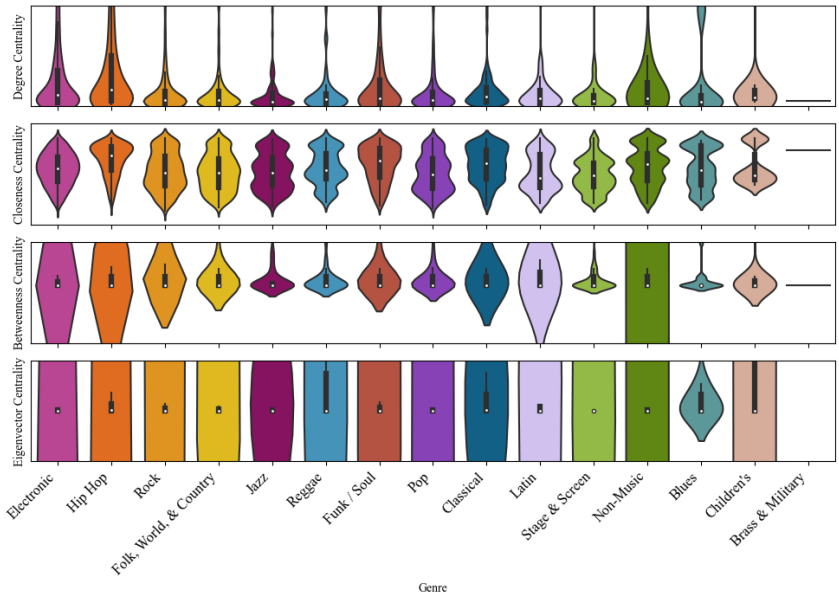


Figure 12: Violinplot of centrality measures vs genre

Figure 12 presents a violin plot that depicts the relationship between centrality measures and genre. The plot's area represents the number of occurrences in each genre, while the boxplot shows the median and interquartile range (IQR), providing insights into the centrality-genre relationship. Appendix 9 displays the corresponding table. Firstly, The general shape of the violin plot of degree centrality illustrates that most values lie around a low median. It is also shown that 'Hip Hop' has the median highest degree of centrality, with a larger inter quartile range (IQR) than other genres, indicating a larger variety in the distribution of the values. 'Classical' and 'Children's' is also higher than other values but with a much smaller range, indicating that most values are relatively close to the median. This concludes that if an artist has the genre 'Hip Hop' assigned, they are more likely to have a higher number of connections in the network.

The violin plot for closeness centrality concludes that the distribution of values is more diverse throughout the genres. Brass & Military has the highest median; since only one node in the giant

component is assigned this genre, we can conclude that this single node holds a very central position in terms of connectivity.'Hip Hop' has the second-highest median closeness centrality, with a small IQR indicating that most values are centred around the median. Furthermore, 'Funk / Soul' displays a similar shape with a relatively high value. 'Electronic', 'Rock', and 'Folk, World & Country' display a similar distribution of results with little differences in the median.

The median range for betweenness centrality for all values is 0, indicating that there is no correlation to be found betweenness centrality and any genre. The IQR is the largest for 'Latin' (0.0000 - 0.0005). A high betweenness centrality is not correlated to a specific genre in the network. Outliers in the network largely determine the shape of the violin plot. The eigenvector centrality shows that a number of genres are more likely to a higher eigenvector value, which suggests an connection to other influential artists.

Overall 'Hip Hop' overall scores well for most centrality measures, indicating that artists attributed with this genre often are more influential within the network. For the other genres, it depends on the measure how well they perform. The more the genre is to the right of the plot the smaller the subset of artists within is, thus outliers have an higher impact on the results.
Figure 13 illustrates the relationship between centrality measures and style in a violin plot. The same settings were used as Figure 12 configuration. The colour of the violin plot represents the genre they were categorised in, as style is considered lower in the hierarchical classification of music. Only the 15 largest styles were shown as Section 5.2 revealed that those cover 41.9% of the network. Notably, the shape of the violin plot mostly follows the same outline as defined by the genre, which is to be expected as the nodes in style that fall under the same genre are the ones that form the violin plot for the genre in Figure 12. Also, the observations made in Figure 12 largely correspond with what would be noticed here.

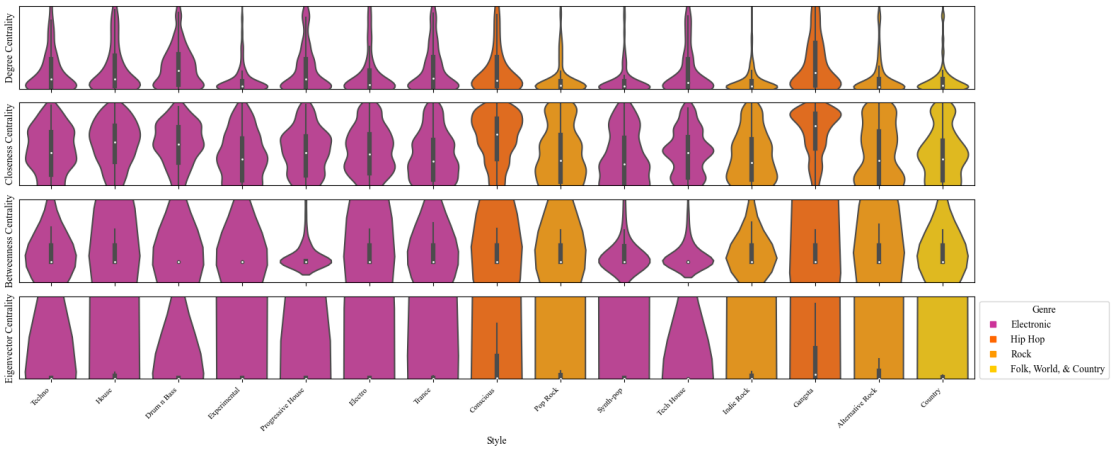


Figure 13: Violinplot of centrality measures vs style

5.3.3 Bridges in the network

In Section 4.5.5, we discussed the use of betweenness centrality to identify entities in the network that act as key players through their role as bridges. Table 2 presented the top five artists with the highest betweenness centrality in the network. In order to assess the impact of removing these influential nodes, we examined the changes in the number of components and the average path length in Table 3.

As depicted in Table 3, the removal of the node with the highest betweenness centrality resulted in the network breaking into 35 separate components. Further removal of nodes did not significantly increase the fragmentation of the network. However, the presence of a large number of nodes in the giant component suggests that the removal of nodes mainly led to the formation of smaller components consisting primarily of single nodes or small groups. Thus, the remaining network remained connected well enough, and the removal of nodes did not cause substantial fragmentation. In conclusion, the average path length in the giant component showed a slight increase as more nodes were removed. This indicates that removing a small group of key players, while slightly increasing the average path length did not result in greater fragmentation beyond small groups of isolated nodes. The role of bridges to break up the network is shown to be minimal, however this does not rule out their function of connecting different groups of artists to each other.

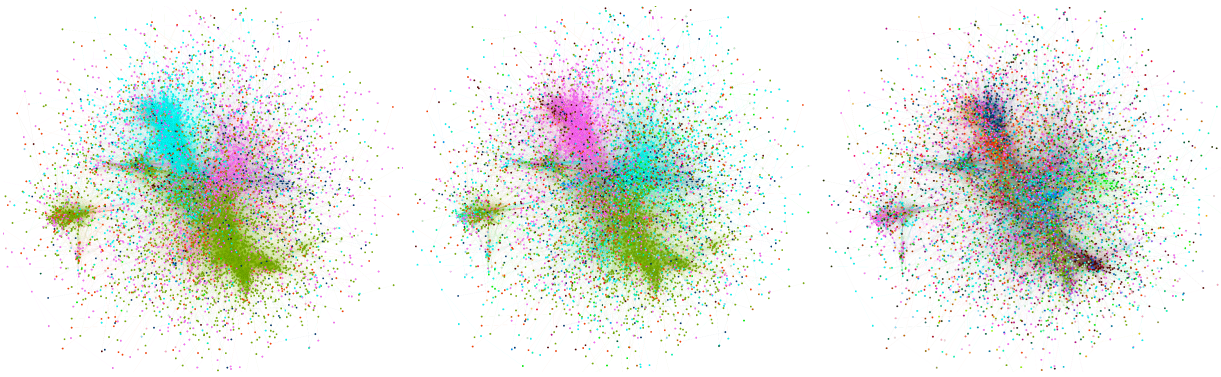
Top # nodes removed	#Components	#Nodes	#Edges	Average path length
0	1	8151	124,058	0.276
1	35	8112	123,407	0.285
5	37	8104	122,539	0.288
10	39	8097	121,160	0.292
20	47	8076	119,024	0.300

Table 3: Removal of bridges (nodes with high betweenness centrality)

5.4 Community structure

To gain insights into the relationships between artists in the network, we examine the network at a meso level by focusing on groups and communities. We investigate whether these communities consist of artists who share common attributes. This analysis allows us to determine whether the community structure in the network is influenced by domain-specific data.

An initial visualisation of the network allows for identifying potential patterns in the network that could indicate the clustering of nodes that share similar attributes. Figure 14 illustrates the network, where the nodes are coloured by attribute value. The attributes showcased are from left to right: ‘Genre’, ‘Refined genre’ and ‘style’. The attribute ‘Label’ contained too many distinct values, that clustering of values could not be seen on this scale. Figure 14a and Figure 14b show some clustering, where the colours are mainly grouped together in three main groups. This could indicate that artists who share a similar ‘Genre’ or ‘Genre refined’ value are more likely to be grouped together. Figure 14c also showcases clustering. However, due to the high number of possible values, the clusters are smaller and less discernible at this scale.



(a) Genre (b) Refined genre (c) Style

Figure 14: artists' co-follower network visualization with nodes coloured by attribute values

5.4.1 Artist with common attributes

This section presents the results of our analysis focusing on attribute-based communities. We used two measures to examine this: the attribute assortativity coefficient and affinity. Both measures calculate the tendency for two artists with the same attribute to be connected. The distinction between them is that the assortativity coefficient focuses on the current network and the proportions of attribute values. On the other hand, affinity compares the proportions of attributes to those expected in a random network, providing insights into the degree of likelihood for attribute values to be connected more than by chance.

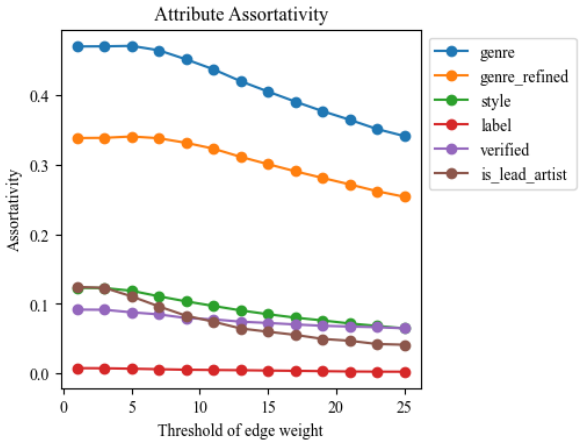


Figure 15: Attribute assortativity on thresholded edge weights

Firstly, we experiment with calculating the attribute assortativity in the network for both music-related attributes and Twitter-related attributes. Since this measure is unweighted we test with different weight thresholds to see if the low weight edges potentially obfuscate the results. Figure 15 displays the attribute assortativity in the network for different weight thresholds. The assortativity coefficient decreases for all attributes as the weight threshold increases. This suggests that high-weight links in the network do not solely exist because of their attributes. The low-weight

links in the network show to be important in the preference of nodes to attach to others that are similar. Furthermore, all attributes showcase a non-negative assortativity coefficient. The assortativity coefficient for 'label' displays a minor positive correlation, while 'genre' seems to display a moderate positive assortativity in the network. The 'refined genre' has a lower value than the 'genre', indicating no higher preference for nodes within the newly introduced genres to be connected to each other. The attributes, 'verified' and 'is_lead_artist', both display a slight positive correlation, suggesting that the preference for a node to connect with someone verified or a lead artist is only slightly stronger.

Secondly, we examine the affinity of an attribute as described in 4. Table 4 displays the affinity values for the attributes in the network. It is observed that all attributes have a significant affinity (affinity > 1), indicating that an artist's connections in the artists' co-follower network are not random but related to their attributes. The affinity value is higher for certain attributes than for others. For example, the release label of an artist is 13.76, implying that artists are 13 times more likely to share the same release label as expected in a random graph. This indicates a strong preference for artists with the same release label to be connected. Similarly, other attributes such as genre, refined genres and style also exhibit significantly higher affinity values in varying degrees. This indicates that artists in the network are more likely to share followers with artists that carry similar attributes than would be the case in a random graph.

Attribute	Affinity
Genre	2.73
Refined Genres	3.04
Style	8.07
Label	13.76
Verified	1.03
Lead artist	1.17

Table 4: Affinity values for attributes in network

Consequently, by combining the results of the attribute assortativity coefficient and affinity measurement, we gain a deeper understanding of the presence of attributes in the network and their tendency to connect artists with similar attributes. The attribute assortativity coefficient for the 'label' attribute is extremely low, indicating a low preference for connecting with artists who share the same label. However, even with this low preference, the likelihood of connecting with artists with the same label is still 13 times higher than in a random network.For the 'genre' attribute, the attribute assortativity coefficient shows moderate positive assortativity in the network. This suggests that artists with similar genres are more likely to be connected, although the likelihood is only about three times higher compared to a random graph.

These measurements provide insights into the connectivity patterns of attributes in the network and allow us to compare them to a random graph with a similar degree distribution. From this analysis, we can conclude that artists are more likely to connect with others who have similar attributes. This suggests that Twitter users are inclined to follow artists who share these attributes.

5.4.2 Attribute-based communities

In Section 5.4.1, we discovered a positive correlation between the attributes of artists and their tendency to connect with other artists with the same attributes. However, the observation that this relationship exists does not guarantee that these nodes are from dense clusters within the network. To investigate whether communities based on common attributes exist in our network, we use modularity to compute the quality of the graph partitioning.

Table 5 presents the modularity for artists when partitioning on the music-related attributes of artists (genre, refined genres, style and label). Additionally, we examined the modularity of partitions created by combining multiple attributes. The results show that the highest modularity (0.233) is found when partitioning the network by genre and a modularity of 0.177 when dividing by refined genres. The rest of the partitions yield a value close to 0, indicating no community structure exists when partitioning the network on these attributes. The partitioning over any set that includes style and label even values to negative modularity indicates that the connections between the artist with common attributes are completely random or worse than random. The large number of communities found, when including the label, over-segments the data indicating a poor fit. It is worth noting that while the findings partially confirm the effectiveness of grouping based on the attribute 'genre,' the results suggest that the traditional methods of grouping artists using other attributes do not exhibit a strong community structure within the network.

Attribute	Communities	Modularity
Genre, Style, Label	6765	-0.0002
Refined genres, Style, Label	6959	-0.0009
Genre, Style	641	0.056
Refined genres, Style	863	0.044
Genre, Label	5772	0.001
Refined genres, Label	6090	0.0002
Style, Label	6633	-0.00006
Genre	15	0.223
Refined genres	21	0.177
Style	298	0.060
Label	5351	0.002

Table 5: Modularity values for communities of musical attributes of artists

5.4.3 Community detection

In addition to examining attribute-based communities, we also conducted community detection experiments to detect groups of nodes that are more densely interconnected with each other. Table 6 presents the modularity for different community detection algorithms and the number of communities they formed. Louvain algorithm yields the highest modularity, and Leiden and InfoMap score close in the modularity range as well, indicating a moderately strong partitioning of the network. Leiden algorithm identified 2641 communities, indicating that many small communities are formed. Almost all community detection algorithms yield a higher modularity value than the

modularity values for attribute-based communities in Table 5.

Algorithm	Communities	Modularity
Louvain	29	0.677
Louvain (res 5)	120	0.356
Louvain (res 35)	659	0.122
Leiden	2641	0.625
InfoMap	452	0.649

Table 6: Modularity values for community detection algorithms

To evaluate the quality of the detected communities with the music-related attributes, we employed the *Normalised Mutual Information* (NMI), a measure explained in detail in Appendix B. Figure 16 visually presents the NMI scores for the different community detection algorithms when compared to the attributes of artists. Notably, the Louvain algorithm scores the lowest for all attributes, indicating a weaker alignment between its detected communities and the genres of the artists. In contrast, both the InfoMap and Leiden algorithms demonstrate higher NMI scores, suggesting a better alignment between their communities and the genres of artists.

The community detection algorithm chosen to continue analysis with is InfoMap based on the results from Table 6 and Figure 16. While Louvain algorithm and Leiden algorithm both score well in modularity, the Leiden algorithm partitions into many communities and the Louvain algorithm does not yield as high score when in the NMI experiment as InfoMap.

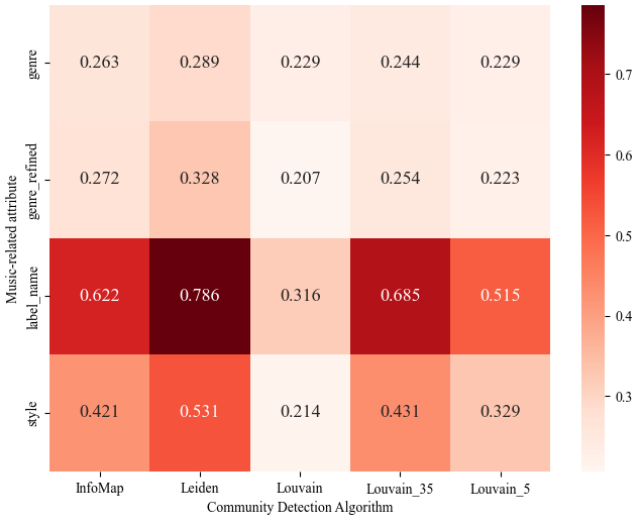


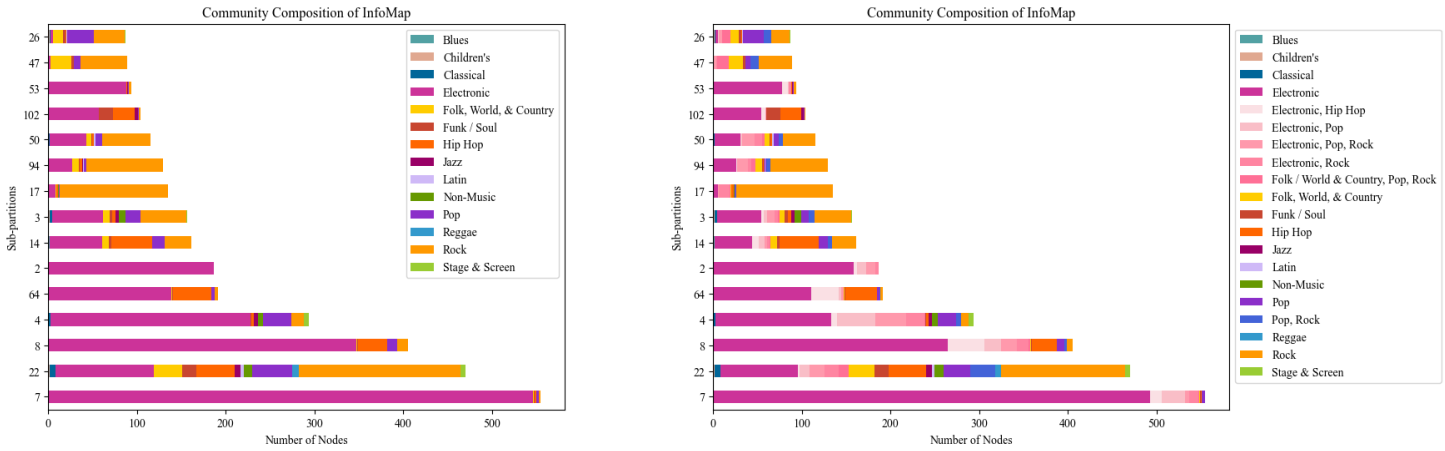
Figure 16: NMI: comparison of community detection algorithms with music-related attributes

After observing a moderate likelihood of artists with the same 'genre' having overlapping followers, we decided to explore how well the communities detected by the InfoMap algorithm align with different 'genre' values. Figure 17 reveals for the Infomap algorithm the communities that were

detected and what 'genre' value they consisted out of. The figures showcase the 15 largest communities. Figure 17 illustrates the composition of the detected communities, specifically showing the 15 largest communities. Figure 17a focuses on the composition of sub-partitions based on 'genre'. Notably, the six largest sub-partitions prominently feature the 'Electronic' genre, except for the second-largest sub-partition. Appendix D displays the same analysis for the other community detection algorithms.

The distribution of the same genres across multiple communities raises the question whether there was more variability within a single genre than initially anticipated. To explore this, we introduced six new refined genres and examine their distribution in the sub-partitions, as depicted in Figure 17b. The overall partition structure remains largely similar, although the presence of 'Electronic' decreases, making room for the newly introduced genres. However, contrary to our expectation, the results indicate that the refined genres not form distinct partitions on their own. Instead, they appear scattered across the sub-partitions identified by the InfoMap algorithm. This suggests that refining the genres does not result in separate artist communities of multiple genres that are more tightly interconnected with each other.

However, upon closer examination, it becomes evident that most tightly-knit clusters in the network consist of a mix of genres, albeit with varying degrees across sub-partitions. Although the partitions primarily revolve around one dominant genre, there are smaller sets of other genres present. This indicates a diverse audience overlap among artists, suggesting that Twitter users predominantly follow artists representing the same genre but not exclusively.



(a) InfoMap community partition by genre

(b) InfoMap community partition by refined genres

Figure 17: Composition of attributes: genre and refined genres in community partition based on InfoMap algorithm

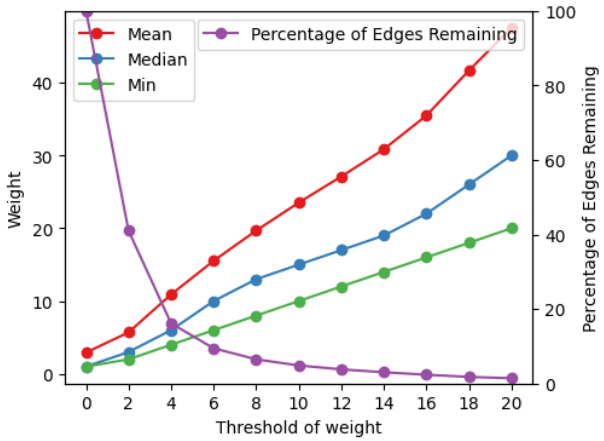
5.5 Robustness analysis

To assess if the chosen pruning method from Section 4.3 does not influence the results derived from the network too much, we conducted a robustness check by constructing an alternative network using edge weight-based pruning instead of the Marginal Likelihood Filter described in Section 4.3. This approach aimed to evaluate the influence of thresholding on significance versus weight on our network analysis results.

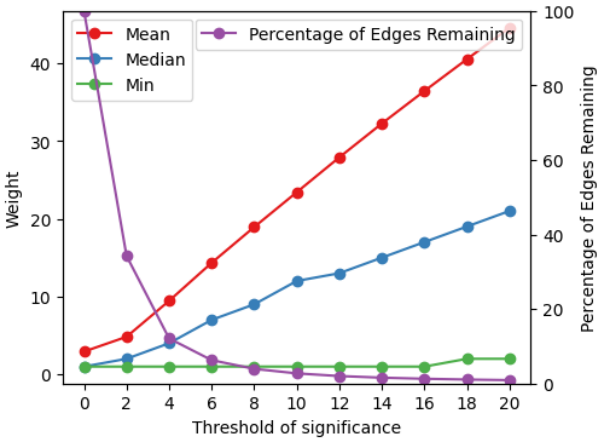
Firstly, we conducted a comparison of statistical measures to examine how they change when pruning the original network using both weight and significance thresholds. Figure 18 illustrates the different changes in these measures based on the two pruning methods. The mean weight for both methods increases at the same rate as the number of remaining edges decreases. In weight thresholding, as shown in Figure 18a, this is because the edges with the lowest weight are removed, resulting in a network where only higher weight edges remain. This is also evident in the rising of the lowest value (min) in the network, which is equal to the threshold. On the other hand, for thresholding by significance, depicted in Figure 18b, the lowest value does not increase in weight until a significance threshold of 16 is reached. To maintain the increase in mean weight, the significance threshold also prunes edges that are deemed insignificant despite having a high weight value. These observations are further supported by the median value. In weight thresholding, the median rises more significantly because only higher weight values remain. However, in significance thresholding, the median stays lower due to the removal of both low and high weight edges from the network. Since the removal rate of low weight edges is higher than that of high weight edges (as indicated by the positive slope of the line in Figure 18b), the median remains relatively lower. These findings suggest that the significance thresholding used for the artists' co-follower network do follow the same statistical measures to some extend, except for the lowest value in the network. The artists' co-follower network disadvantages artists with a smaller following to a lesser extend than with pruning by weight.

We filter approximately the same number of edges (98.69%) by applying a threshold that removes all edges with a weight ≤ 21. Firstly, a comparison is drawn between the network measures of the artists' co-follower network and this network. Table 7 shows the network measures for a weight threshold of 21. The number of nodes in the giant component is only a quarter of all the nodes present in the artists' co-follower network, when removing the same number of edges. This indicates that many more nodes become isolates when pruning the same number of edges with weight than the MLF filter does. This is expected as the MLF filter aims to extract a sparser graph containing more nodes. The average degree of the nodes is 66.55, which is almost 4 times higher than reported in our network. This is not unexpected, as with roughly the same number of edges; fewer nodes are to be connected.

Comparing the degree and weight distributions in the network provides insights into the network's structure and relationship strength. Figure 19 illustrates the degree distribution using two thresholding methods. The weight threshold method (Figure 19a) results in higher overall degrees. This aligns with the findings in Table 7, where the same number of edges is distributed among fewer nodes, resulting in higher degrees per node compared to the artists' co-follower network (Figure 19b). Both methods follow a similar shape, indicating a power-law distribution in the network's degree



(a) Mean, median, min, and percentage of edges vs weight threshold



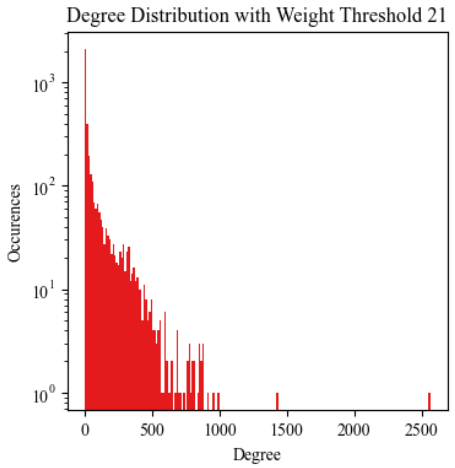
(b) Mean, median, min, and percentage of edges vs significance threshold

Figure 18: Comparison of statistical measures when applying different network pruning methods

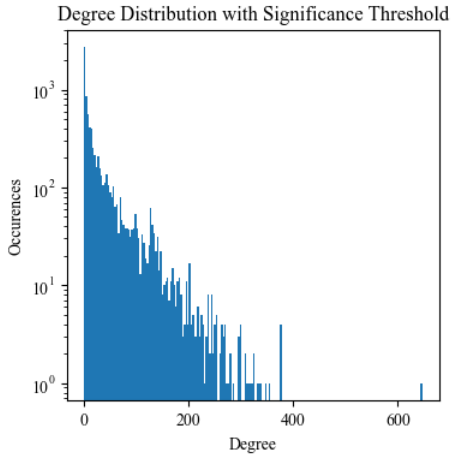
Network measure	**Values**
Number of nodes	3808
Number of edges	126,710
Average degree	66.55

Table 7: Network statistics of the artists' co-follower network with weight threshold 21

for both measures.



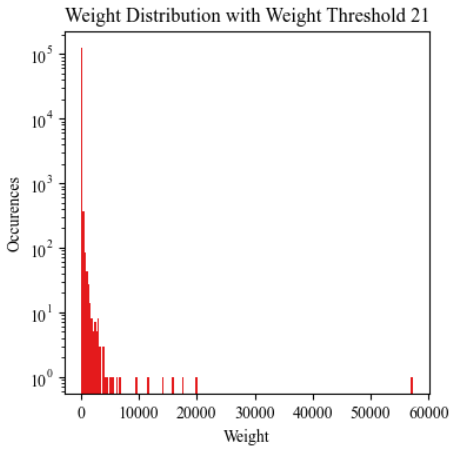
(a) Degree distribution for weight threshold



(b) Degree distribution for significance threshold

Figure 19: Degree distribution for weight and significance thresholding on the network

In addition to that, we also compare the weight distribution to see how robust the network is. Figure 20 depicts the weight distribution for both pruning methods. The distribution follows largely the same shape indicating that the method of pruning does not have a significant effect on the weight distribution in the network.



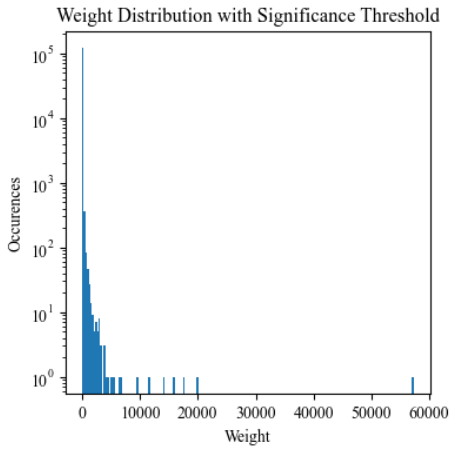


Figure 20: The weight distribution for weight and significance thresholding on the network

The robustness analysis compared weight and significance thresholding methods on a network. While weight thresholding resulted in higher mean weight, more isolates, and higher average degree, both methods exhibited similar degree and weight distributions, suggesting that the overall structure and relationship strength of the network were not significantly affected by the choice of pruning method.

6 Conclusion and future directions

In this thesis, we introduced a method for analysing the artists' co-follower network on Twitter. Through the use of social network analysis, we made an effort to understand this network and its implications. Data collection was performed using various data sources, such as Discogs for a list of artists and a Twitter snapshot for the relationships. Through a process of data cleaning, a link was created between these sources to obtain the artists' co-follower network. We identified a hierarchy property in the network with a low average clustering coefficient but a high number of triangles suggesting that a small group of nodes is highly connected. The network also inhabited the property of power law for both degree and edge weight. Also, we managed to roughly get the same distribution of genres throughout the network. However, some genres were overrepresented by the limitation of attribute aggregation. We made an effort to mitigate this by introducing a refined set of genres. While these lowered the over-representation slightly, we could not entirely remove this. We found the most influential artist in the network and identified that the reach of the 5 artists with the highest degree of centrality largely followed the same group of artists.

In addition, we found that the set of artists linked to the top 5 artists with degree centrality has an average degree that is 350% higher than average. This further supports the idea of hubs in this network. We investigated if the relationships found in centrality between the biggest artists propagated to the rest of the network. Through correlation analysis, we revealed a slight correlation between closeness centrality and degree centrality and one between eigenvector centrality and betweenness centrality. We also uncovered that artists with a higher eigenvector centrality and betweenness centrality are slightly more likely to have a higher follower count. In addition, we found that removing artists with high betweenness centrality does not cause the network to become fragmented into large disconnected components. Also, the relationship between the musical attributes of the artist and the centrality measures indicates that the attribute 'style' largely follows the same shape as the genre attribute that is assigned to artists, in addition to that finding relationships between attributes and centrality measures was difficult but for 'Hip Hop' it was shown they on average rank higher for all centrality measures.

We showed that attribute-based communities exist in the network but do not yield high modularity scores. However, we also showed that most attributes are more connected in this network than in a random network, indicating some preference for artists with similar attributes to have an overlap in followers. This shows us that homophily is indeed applicable to this specific network. The community detection algorithm shows that the partitions found do not necessarily align with defined genres but are somewhat splintered across the sub-partitions. Even the introduction of a refined genre set did not create sub-partitions solely based on one genre. This indicates that the audience overlap between artists is relatively diverse in their following.
Summarising the work, we have researched artists, their relations, and relationships between music-related attributes and network measures such as centrality and community detection. Also, we investigated traditional community assignments and how they translate to community detection algorithms. These results can help understand artists' positions in the network and their similarity in audience with other artists, which can be helpful to look into promotion or collaborations in the music industry and reach new and diverse audiences.

6.1 Future research

One main limitation of our work is that the Twitter snapshot data is not up to date. For future research, incorporating a more recent dataset would enable a more comprehensive network analysis of artists in the present time. Additionally, gathering information about the Twitter users who follow artists could provide new insights into the characteristics of the audiences that follow the same artists, shedding light on the clustering patterns observed in this research. Furthermore, conducting temporal analysis would offer valuable insights into the evolving dynamics of music culture. Considering that the popularity of genres changes over time, it would be interesting to investigate whether these changes are reflected in the influential artists and communities detected within the network. This temporal perspective would contribute to a better understanding of how music culture propagates online over time and its impact on the network structure.
Overall, incorporating a more recent dataset, exploring the characteristics of artist followers, and conducting temporal analysis would enhance our understanding of the evolving music culture and its manifestation in online networks.

References

[1] Matplotlib; Visualization with Python — matplotlib.org. https://matplotlib.org. February, 2023.

[2] NetworkX; NetworkX documentation — networkx.org. https://networkx.org. March 5th, 2023.

[3] Seaborn: statistical data visualization; seaborn 0.12.2 documentation — seaborn.pydata.org. https://seaborn.pydata.org. [April, 2023].

[4] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10):P10008, October 2008.

[5] S. Borgatti. Identifying sets of key players in a social network. *Computational & Mathematical Organization Theory*, 12:21–34, April 2006.

[6] N. Dianati. Unwinding the hairball graph: Pruning algorithms for weighted complex networks. *Physical Review E*, 93(1), January 2016.

[7] L. C. Freeman. *The development of social network analysis a study in the sociology of Science*, page Chapter 1: Introduction. Empirical Press, 2004.

[8] K. Garrison, D. Scheinost, E. Finn, X. Shen, and R. Constable. The (in)stability of functional brain network measures across thresholds. *NeuroImage*, 118, May 2015.

[9] J. Golbeck. *Analyzing the Social Web*. Morgan Kaufmann, San Francisco, 2013.

[10] J. Hartnett. Discogs.com. *The Charleston Advisor*, 16(4):26–33, April 2015.

[11] I. Himelboim. *Social Network Analysis (Social Media)*. August 2017.

[12] H. Kwak, C. Lee, H. Park, and S. Moon. What is twitter, a social network or a news media? In *Proceedings of the 19th International World Wide Web Conference (WWW '10)*, pages 591–600, 2010.

[13] R. Lambiotte and M. Ausloos. On the genre-fication of music: A percolation approach. *The European Physical Journal B - Condensed Matter and Complex Systems*, 50(1–2):183–188, April 2006.

[14] A. Lancichinetti, S. Fortunato, and J. Kertész. Detecting the overlapping and hierarchical community structure in complex networks. *New Journal of Physics*, 11(3):033015, mar 2009.

[15] J. Leskovec and A. Krevl. SNAP Datasets: Stanford large network dataset collection. https://snap.stanford.edu/data/twitter-2010.html, June 2014.

[16] S. McAndrew and M. Everett. Music as Collective Invention: A Social Network Analysis of Composers. *Cultural sociology*, 9(1):56–80, 2015.

[17] A. McDaid, D. Greene, and N. Hurley. Normalized mutual information to evaluate overlapping community finding algorithms. *CoRR*, October 2011.

[18] M. Mcpherson, L. Smith-Lovin, and J. Cook. Birds of a feather: Homophily in social networks. *Annual Review of Sociology*, 27:415–, January 2001.

[19] A. Mislove, M. Marcon, K. P. Gummadi, P. Druschel, and B. Bhattacharjee. *Measurement and Analysis of Online Social Networks*. PhD thesis, October 2007.

[20] M. E. J. Newman. Mixing patterns in networks. 67(2), feb 2003.

[21] T. Opsahl. Triadic closure in two-mode networks: Redefining the global and local clustering coefficients. *Social Networks*, 35(2):159–167, 2013. Special Issue on Advances in Two-mode Social Networks.

[22] T. Opsahl, F. Agneessens, and J. Skvoretz. Node centrality in weighted networks: Generalizing degree and shortest paths. *Social Networks*, 32(3):245–251, 2010.

[23] M. Rosvall, D. Axelsson, and C. T. Bergstrom. The map equation. *The European physical journal. ST, Special topics*, 178(1):13–23, 2009.

[24] A. Saxena and S. Iyengar. Centrality measures in complex networks: A survey. *arXiv preprint arXiv:2011.07190*, 2020.

[25] V. A. Traag, L. Waltman, and N. J. van Eck. From louvain to leiden: guaranteeing well-connected communities. *Scientific Reports*, 9(1), mar 2019.

[26] J. Vlegels and J. Lievens. Music classification, genres, and taste patterns: A ground-up network analysis on the clustering of artist preferences. *Poetics*, 60, September 2016.

[27] H. Wi, K. H. Hyun, J. Lee, and W. Lee. The effect of djs' social network on music popularity. 01 2016.

[28] T. Zhou, J. Ren, M. Medo, and Y.-C. Zhang. Bipartite network projection and personal recommendation. *Physical review. E, Statistical, nonlinear, and soft matter physics*, 76:046115, 11 2007.

A Attribute aggregation

Listing 1: Attribute aggregation based on most occurences

```
-- Genre, label can be found by changing "style" to "genre"/"label_name"
-- and the respective table

with tog AS (
SELECT sanlielwa.artist_id, ra.release_id, rg.style
FROM stan_artist_node_list_in_edge_list_with_attributes sanlielwa
LEFT JOIN release_artist ra
        ON sanlielwa.artist_id = ra.artist_id
LEFT JOIN release_style rg
        ON ra.release_id = rg.release_id
),
find_style AS (SELECT artist_id, style, count(style) AS style_count
FROM tog
WHERE style IS NOT NULL
GROUP BY artist_id, style
ORDER BY artist_id, style_count DESC, style
),
ranking AS (SELECT *, row_number() over (PARTITION BY artist_id
ORDER BY style_count desc) AS rank
FROM find_style
),
style_per_artist AS (SELECT artist_id, style
FROM ranking
WHERE rank = 1
)
UPDATE stan_artist_node_list_in_edge_list_with_attributes AS san
SET style = gr.style
FROM style_per_artist as gr
WHERE san.artist_id = gr.artist_id
```

B (Normalized) mutual information

Mutual information is a quantitative measure used to compare different data partitions and assess their similarity. This method has gained popularity in social network analysis, particularly for evaluating community detection algorithms that identify overlapping clusters of nodes [17]. When considering two cluster partitions, denoted as X and Y, mutual information represents the intersection of nodes belonging to both partitions. Figure 21 provides an illustration of partitions X and Y, with $I(X:Y)$ representing the shared nodes between them.
Normalisation is applied by division by a value that is bound to be the upper bound. Creating a range of values between 0 and 1 [17]. Normalized mutual information [14] is defined as:

$$I(X:Y) = \frac{H(X) + H(Y) - H(X,Y)}{(H(X) + H(Y))/2} \tag{3}$$

Where the entropy of the random variable X and Y is denoted as respectively $H(X)$ and $H(Y)$ in association with their partition, $H(X,Y)$ is the joint entropy of both variables. Because of normalisation, the variable is in the range $[0,1]$ and values 1 when the two partitions are entirely equal and 0 when they are entirely unequal.

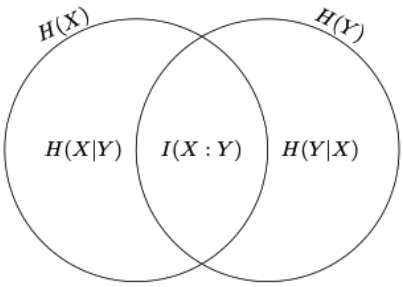


Figure 21: Venn diagram of mutual information [17]

C Centrality

	Degree Centrality	Closeness Centrality	Betweenness Centrality	Eigenvector Centrality	Follower Count
Degree Centrality	1	0.605786	0.254785	0.248149	0.076507
Closeness Centrality	0.605786	1	0.076515	0.088895	0.020937
Betweenness Centrality	0.254785	0.076515	1	0.651864	0.466026
Eigenvector Centrality	0.248149	0.088895	0.651864	1	0.496066
Follower Count	0.076507	0.020937	0.466026	0.496066	1

Table 8: Correlation matrix of centrality measures

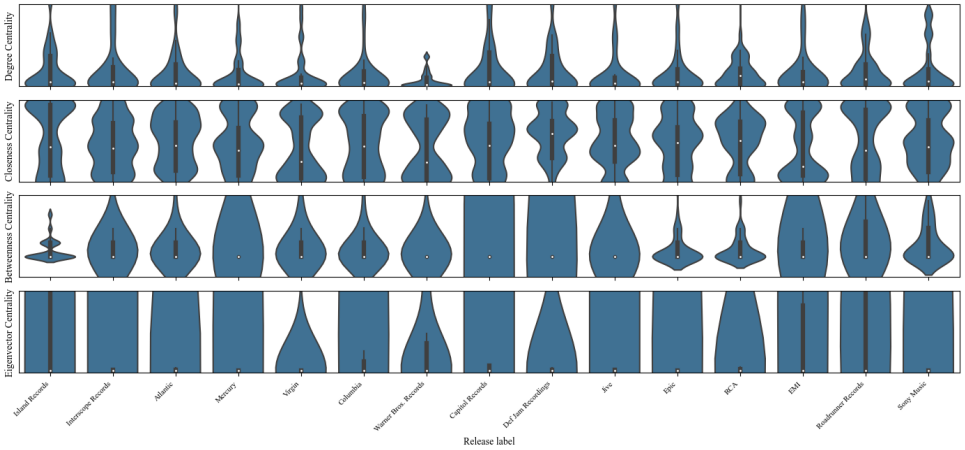


Figure 22: Violinplot of centrality measures vs label

	Degree Centrality		Closeness Centrality		Betweenness Centrality		Eigenvector Centrality	
	Median	Range (IQR)	Median	Range (IQR)	Median	Range (IQR)	Median	Range (IQR)
Electronic	0.019	0.003 - 0.068	0.577	0.401 - 0.740	0.0000	0.0000 - 0.0001	9.352e-08	2.051e-09 - 1.847e-06
Hip Hop	0.029	0.006 - 0.098	0.759	0.554 - 0.875	0.0000	0.0000 - 0.0003	2.234e-06	5.417e-08 - 2.863e-05
Rock	0.008	0.002 - 0.026	0.520	0.342 - 0.746	0.0000	0.0000 - 0.0003	2.316e-07	1.980e-09 - 1.099e-05
Folk, World & Country	0.008	0.002 - 0.025	0.522	0.318 - 0.713	0.0000	0.0000 - 0.0003	1.227e-07	1.277e-09 - 7.348e-06
Jazz	0.005	0.002 - 0.022	0.527	0.350 - 0.726	0.0000	0.0000 - 0.0001	9.940e-08	9.334e-10 - 2.787e-06
Reggae	0.009	0.002 - 0.020	0.552	0.433 - 0.789	0.0000	0.0000 - 0.0003	4.384e-07	1.038e-09 - 1.470e-04
Funk / Soul	0.011	0.002 - 0.049	0.690	0.459 - 0.850	0.0000	0.0000 - 0.0003	5.139e-07	3.820e-09 - 1.447e-05
Pop	0.008	0.002 - 0.023	0.497	0.303 - 0.707	0.0000	0.0000 - 0.0003	3.622e-08	1.375e-10 - 2.638e-06
Classical	0.015	0.005 - 0.032	0.647	0.439 - 0.832	0.0000	0.0000 - 0.0003	1.769e-06	2.506e-08 - 6.499e-05
Latin	0.011	0.002 - 0.028	0.451	0.321 - 0.765	0.0000	0.0000 - 0.0005	1.192e-07	3.391e-09 - 1.585e-05
Stage & Screen	0.007	0.003 - 0.017	0.487	0.327 - 0.657	0.0000	0.0000 - 0.0003	2.301e-08	2.062e-10 - 9.093e-07
Non-Music	0.011	0.003 - 0.043	0.641	0.416 - 0.784	0.0000	0.0000 - 0.0003	6.215e-08	9.304e-10 - 2.697e-06
Blues	0.005	0.000 - 0.017	0.557	0.356 - 0.890	0.0000	0.0000 - 0.0000	1.302e-07	1.166e-08 - 6.456e-05
Childeren‘s	0.014	0.012 - 0.027	0.485	0.430 - 0.771	0.0000	0.0000 - 0.0001	1.161e-07	3.241e-09 - 3.699e-03
Brass & Military	0.006	0.006 - 0.006	0.836	0.836 - 0.836	0.0000	0.0000 - 0.0000	4.981e-04	4.981e-04 - 4.981e-04

Table 9: Centrality measures by genre: median and IQR

D Community composition of community detection algorithms

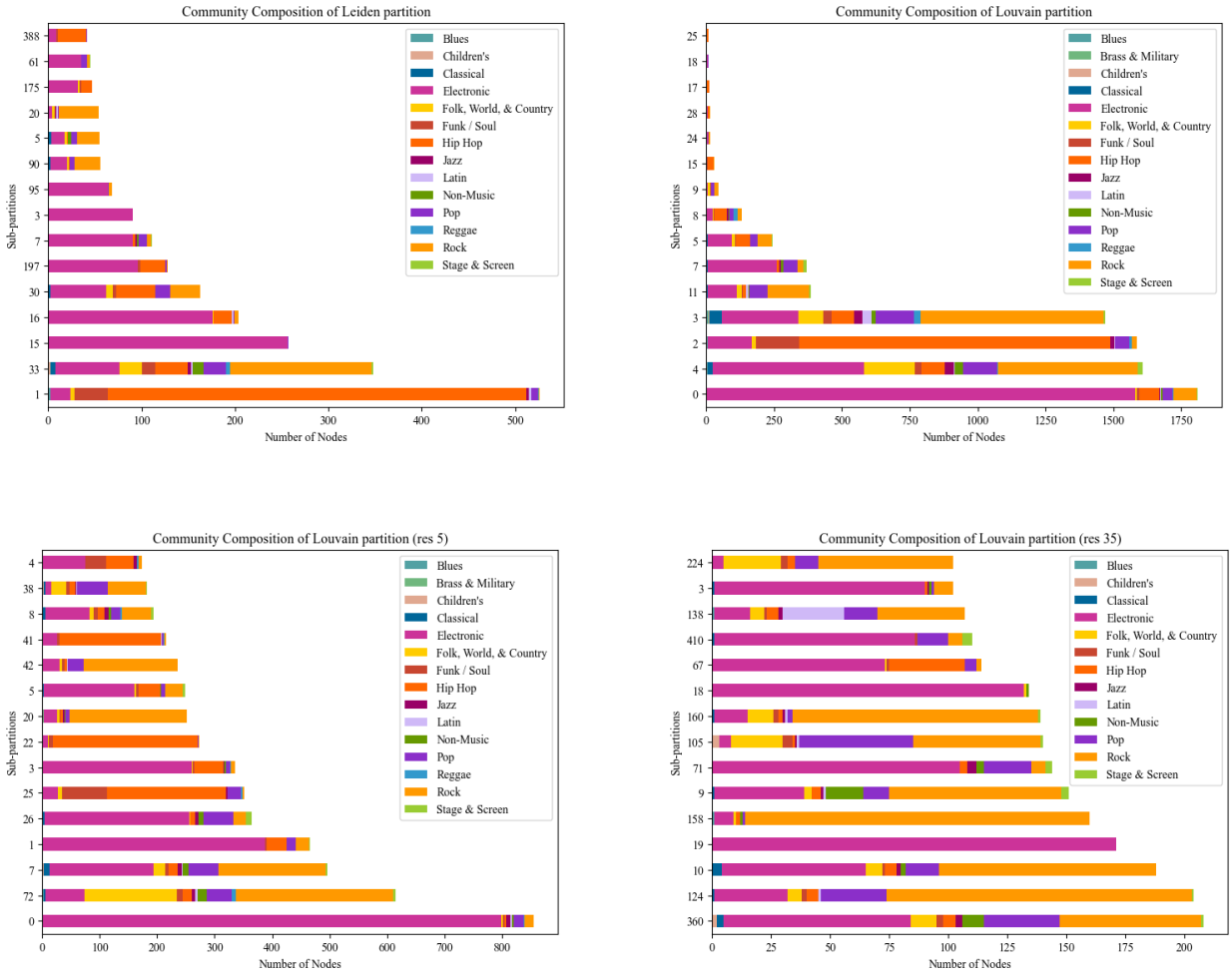


Figure 24: Community composition of community detection algorithms based on genre