Demographic analysis of music preferences in streaming service networks

Lidija Jovanovska[1,2][0000−0001−7430−1015], Bojan Evkoski[1][0000−0003−1096−4666], Miroslav Mirchev[1][0000−0001−9899−2439], and Igor Mishkovski[1][0000−0003−1137−6102]

[1] Faculty of Computer Science and Engineering,
Ss. Cyril and Methodius University, 1000 Skopje, North Macedonia
[2] Jožef Stefan International Postgraduate School, 1000 Ljubljana, Slovenia
lidija.jovanovska@outlook.com

Abstract. As Daniel J. Levitin noted, music is a cross-cultural phenomenon, a ubiquitous activity found in every known human culture. It is indeed, a living matter that flows through cultures, which makes it a complex system potentially holding valuable information. Therefore, we model country-to-country interactions to reveal macro-level music trends. The purpose of this paper is twofold. Firstly, we explore the way specific demographic characteristics, such as language and geographic location affect the global community structure in streaming service networks. Secondly, we examine whether a clear flow of musical trends exists in the world by identifying countries who are prominent leaders on the music streaming charts. The community analysis shows that there is strong support for the first claim. Next, we find that the flow of musical trends is not strongly directional globally, although we were still able to identify prominent leaders and followers within the communities. The obtained results can further lead to the development of more sophisticated music recommendation systems, kindle new cultural studies and bring discoveries in the field of musicology.

Keywords: Music · Community detection · Leader-follower relationship.

1 Introduction

Throughout the previous decade, the world has experienced a drastic change in the way music is listened to. While physical and digital copy sales continue to fall globally, streaming services usage is growing massively. Namely, streaming income increased by 230% between 2013 and 2017 and it continues to do so year after year by almost 25%, marking it as a period of steady development.

In regards to the fluidity of the CD music market, Ferreira and Walfdogel stated that in at least one year of continuity, 31 artists appeared concurrently on the charts in more than 18 countries on the worldwide music trade between 2001 and 2007 [4]. This suggests a high similarity of the regional musical preferences on a global level. However, digital songs are more likely to fall off the chart in the first week than CD songs, indicating a highly volatile market, resulting regularly in heterogeneous charts [5]. These findings prompted us to examine whether this volatile market possesses clear patterns that could be used for further research and applications such as developing music recommendation systems, running cultural studies and discovering more about the nature of today's music.

Our study is organized as follows: Firstly, in Section 2, we provide an overview of some of the most relevant studies on networks and music. Then, in Section 3, we describe the data taken from the global log of streaming habits recorded by Spotify. In the following Section 4, we describe how the music networks were constructed and investigate whether and to what extent communities are influenced by language and geographical distance when looking at the whole dataset, but also at specific genres. In Section 5, we present an approach for detecting leadership in the music network, by measuring the correlation between top charts of each pair of countries in the dataset. Finally, in Section 6 we conclude the paper with a summary of contributions and avenues for future work.

2 Related Work

The field of Network Science provides a framework for modeling interactions between entities to reveal macro-level properties that may not be noticeable at the individual level. The potential of these methods stems from the fact that the creation and transmission of cultural products are essentially network phenomena. Therefore, networks can lead to a new fundamental understanding of the complex nature of culture in music cognition. In the past decade, there has been plentiful research on music preference due to the demand for better recommendation systems in the online world. This has motivated many researchers to try to better understand the connections in music networks across the globe.

Gunaratna and Menezes analyzed a social network of Brazilian musicians to identify the artists responsible for the flow of information in the network. They also demonstrated that the network follows a power-law distribution, with prominent hierarchical characteristics only during the latter decades (1990-2010). This tendency towards a hierarchical structure can be explained as a result of the increasing availability of collaboration tools, i.e. social media [7]. Salganik et al. studied the effect of social influence on a micro-level (success of a song) and the macro-level (market inequality). They did so by creating an artificial "music market" in which a group of participants rated previously unknown songs either with or without knowledge of previous participants' choices. They discovered that social influence increased both inequality and unpredictability of success [10]. Nagy et al. provided a powerful methodology for detecting leader-follower pairs, which was previously applied in the search for the leadership hierarchy present in pigeon flocks [8]. This was further adapted by Lee and Cunningham in their research about the global flow of music on Last.fm [6], which served as an inspiration for this research.

3 Data

Streaming services provide easy access to rich and various content which can be used for thorough analysis of many properties of music, artists and its consumers. In recent times, Spotify has arguably been the most popular streaming platform with a user base of 248 million active users worldwide. Since January 2017, the Spotify Top 200 Charts are open to the public via their API. It allows easy access to song and artist data, but it also provides daily and weekly worldwide streaming top charts. They are available for each country separately, which makes them suitably formatted for our analysis.

Spotify currently provides streaming services to 79 countries, of which 65 have regular weekly Top 200 Charts. Since we were interested in finding the flow of the musical trends in the network, it was necessary to provide data for the longest time span possible (January 2017 - June 2019). Considering that not every country started providing top charts from day one, the filtered dataset ultimately contained 55 countries.

The working dataset contained over 270 billion streams with a median of 1.6 billion and an average of 5 billion streams per country. It included 6810 different artists with over 31000 unique songs, which managed to climb into the Top 200 in at least one region. Each sample contained the song name, artist, Spotify URL of the song (unique), date, country, the streaming count of the song for the specific week and its position.

Table 1. Basic dataset statistics

# Countries	# Songs	# Artists	# Weeks	Total Streams	Min Streams	Max Streams
55	31093	6810	124	273400 mil.	Lux: 93 mil.	US: 69073 mil.

Finally, to be able to undergo genre-specific analysis, we used the genre hierarchy defined in the Free Music Archive (FMA) [3] to group all 1500 different Spotify genres into their corresponding 12 parent genres defined in FMA: blues, country, electronic, folk, hip-hop, classical, jazz, pop, reggae, rock, soul-rnb and indie/experimental (contains every genre that did not fit the rest).

4 Identifying global communities

To gain knowledge about the connections and the structure of the music streaming world in terms of countries, we attempted to find clusters based on language and geographic distance by analyzing the similarity of their top charts.

4.1 Streaming matrices

For the purpose of creating a suitable mathematical representation of a country's streaming history, the data was aggregated into matrices, which we refer to as streaming matrices. These streaming matrices contain every unique song ever to appear on the Spotify top charts as a column, while each row represents one of the 55 countries. Since most of the countries do not have the same set of songs appearing on their charts, the matrices are sparse. The cells showed the total streaming count of a song for a particular country. In mathematical terms, a non-zero entry in the matrix at position i, j is a positive integer, indicating the total number of times the users from country i have streamed the song j. With this representation, each country was a vector of 31000 values (unique songs). Since there was a big difference in the number of streams between countries (the USA with over 7 billion while Peru with only 300 million) min-max normalization on each row separately was necessary.

Table 2. Streaming matrix example

	Africa	**Christmas Lights**	**...**	**Mr. Brightside**	**Shape of You**
Ireland	4112k	810k	...	6032k	15821k
...	...	...	...	...	...
USA	41639k	1823k	...	9336k	351342k

4.2 Agglomerative clustering

The normalized streaming matrices were used to compute the agglomerative clusters and to create dendrograms by using Ward's distance method. By examining Fig. 1, it became clear that there are two major clusters: a Spanish speaking and non-Spanish speaking, which asserts the influence of language on the development of the major clusters. It is also noticeable that the Spanish speaking countries form their cluster much faster than the rest of the world. The language bond is visible in the other cluster too, where there is a strong connection between the UK and Ireland, Germany and Austria, etc. Strong geographical influences are also present. For example, even though Greek and Bulgarian come from different language families, there is a strong musical similarity between Greece and Bulgaria, due to geographical closeness. The same pattern appears with Hungary and the Czech Republic. This repeats to an even larger extent if we analyze country pairs that use similar languages. Finally, it should be noted that the cluster containing France, Italy, Turkey, and Brazil, which shows up rather late in the groupings is not a cluster at all. As we would discover later, these four countries mostly listen to local music and they have largely independent top charts from the rest of the world.

4.3 Community networks

To be able to visualize the clustering, but also apply some other clustering techniques specific to graphs, undirected community networks were generated for

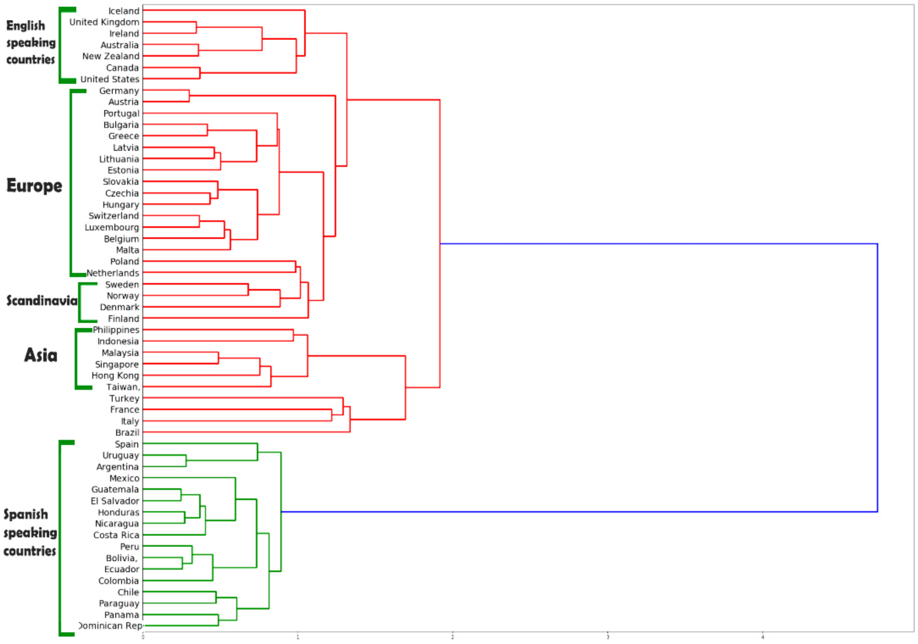

Fig. 1. Agglomerative clustering on top charts similarity

each genre separately. To form edges, the cosine similarity between each vector was measured. Its most frequently used in text analysis, to measure document similarity. However, it can be used in other scenarios which can be represented with vector notation. Afterwards, using Gephi [1], the modularity of the network was calculated [2]. It is a widely used asset in community structure analysis which measures the strength of division of the network into communities.

Fig. 2 shows the community network for the pop music genre where the size of the nodes represents the streaming counts. As expected, the Spanish cluster was present and strongly interconnected, but the other smaller communities revealed much more interesting results. The next smaller clusters were eminent: the English-speaking countries, the Balkan community, Central Europe, Eastern Europe and finally, the Asian countries. Italy and Brazil remained strongly disconnected from the rest of the world, while Paraguay and Finland join them as cultures mostly devoted to domestic music. Furthermore, even though the Spanish language dominates politics and diplomacy in Paraguay, their native Guarani language is used by over 90% of the population. On the other hand, Finland's disconnectedness is also due to the uniqueness of the Finnish language as part of the Uralic language family, unlike the rest of the Scandinavian countries, who are part of the Germanic language family.

The main support of the hypothesis that language has the biggest influence on the development of musical communities is the fact that when looking at the

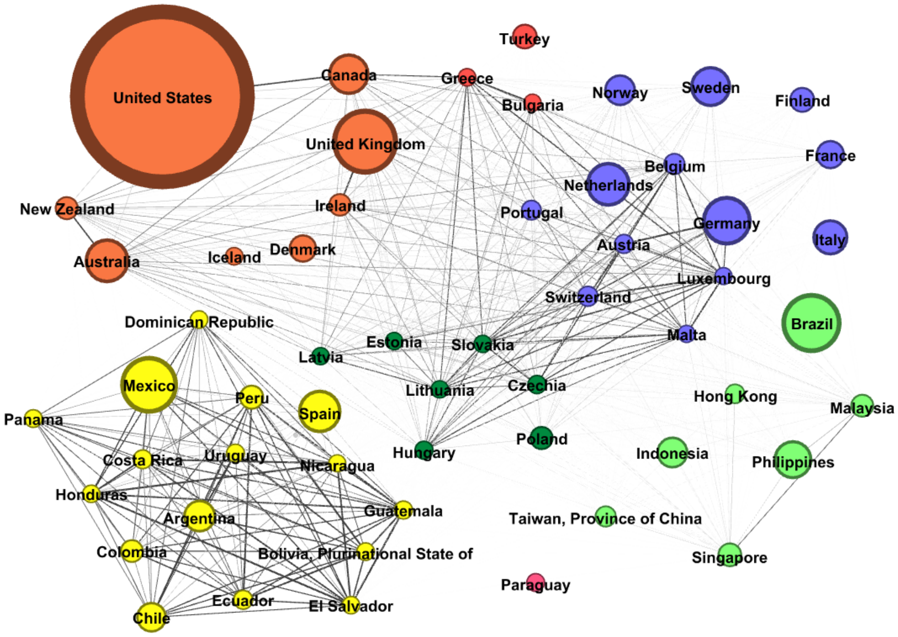

Fig. 2. Pop music top charts community network using modularity score (node size - total stream count, edge weight - cosine similarity)

genre which has almost no words at all - the classical genre, the communities that emerge are significantly more versatile in terms of language and geographical distance (Fig. 3).

5 Detecting global music leaders

The next step was to delve deeper into the characteristics and the dynamics of the obtained network. In the previous section we identified the biggest communities in the network, which can be very helpful for reasons stated above, but detecting the leaders in those communities is another important aspect of acquiring knowledge about the structure of the network. Generally, leaders are considered as countries that dictate the top charts and initiate the trends, while followers are countries that adopt certain trends within a certain time lag. This led to the design of a novel approach for finding leadership in the community network.

5.1 Methodology

When looking at a pair of countries we want to find out whether country i follows country j, j follows i, or no connection exists. Since we have weekly top charts for

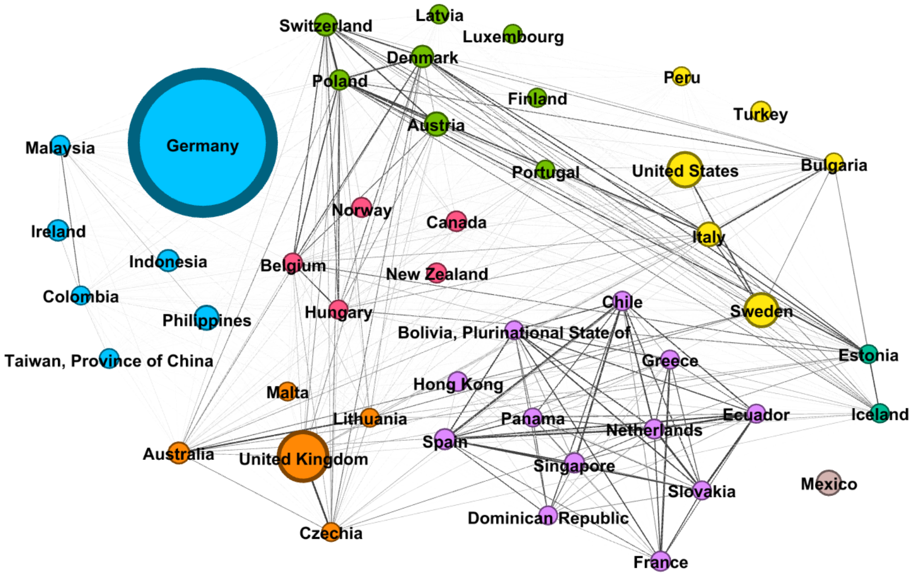

Fig. 3. Classical music top charts community network using modularity score

every country, we determine the relationship by comparing charts of the pair of countries, but in different time steps. For example, if we want to determine the relationship between Sweden and Denmark, we measure the Jaccard similarity of Sweden's top charts in week t, with Denmark's top charts in week t - *1*, and vice versa for every t in the dataset. If the similarity is strong enough, we choose the higher similarity and add a directed link from the leader to the follower, e.g. from Denmark to Sweden. Hence, a simple follower score can be defined as:

$$\text{Follower\_Score}(C_i, C_j) = \sum_t \text{Jaccard}(C_i, C_j) = \sum_t \frac{C_i(t) \cap C_j(t-1)}{C_i(t) \cup C_j(t-1)}.$$

However, to make this approach more reliable, these computations are calculated for different time lags: from one week (as in the example above), to 12 weeks for each pair of countries. We then compute the global averages for every time lag separately, which we take as thresholds that help us determine the significance of the similarities. Afterward, we sum the 12 similarities for each pair and measure how many of those similarities are above their respective threshold, labeled as k in our formula. We use these counts as parameters to the similarity sums. Larger k corresponds to a stronger followership, while a smaller value means a less consistent one. These scaled similarity sums are the weight of the network edges. Note that if k is zero, the edge is removed. The follower score which we propose takes the following form:

$$\text{Follower\_Score}(C_i, C_j) = (\sum_{d=1}^{12} \sum_t \frac{(C_i(t) \cap C_j(t-d))}{(C_i(t) \cup C_j(t-d))}) \cdot k.$$

Finally, to identify if, for example, "Sweden follows Denmark" significantly more than "Denmark follows Sweden", we calculate the average difference between A and B and B and A scores for all pairs of countries. Only if this threshold is crossed, we create an edge between Sweden and Denmark, otherwise we conclude that the following is not significant enough, thus there is no leader-follower relationship.

5.2 Results

After the directed network was generated with the method described above, the well-known PageRank algorithm was used to compute the importance of the nodes [9]. The algorithm utilizes the weights of the edges and helps by enabling a better visualization of the network. In that sense, the size of the nodes in the network represents their PageRank score. As in the previous section, the modularity was measured to see if the same communities that emerged in Section 4 would emerge again.

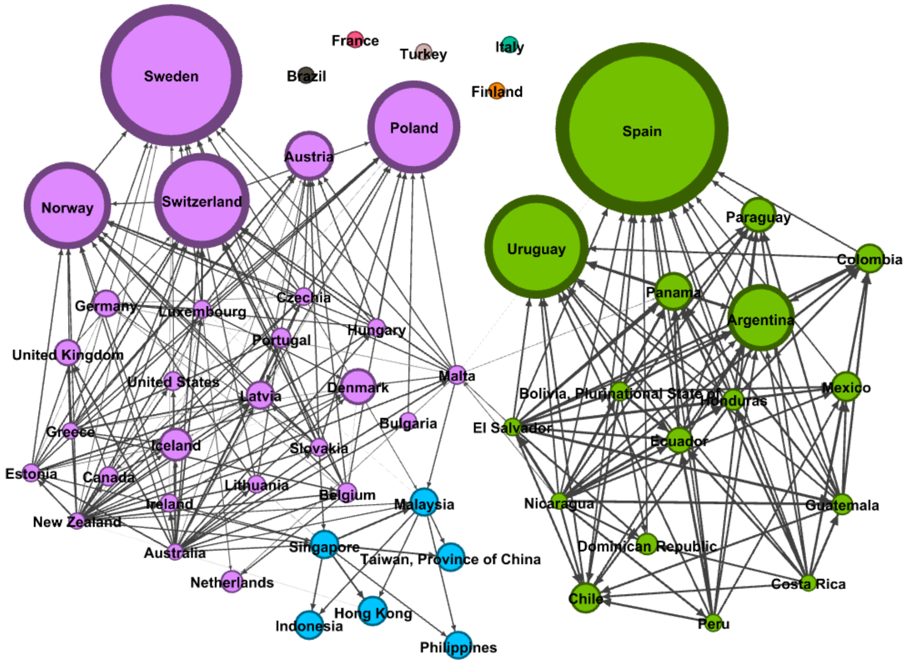

Fig. 4. Music leaders network (node size - PageRank, edge weight - lagged jaccard similarity)

As we can see in Fig. 4, the Spanish speaking countries form a strong cluster once again, with Spain as the obvious leader in this group. Interestingly enough, the other large cluster formed by the non-Spanish Western world is not what we expected to see. Even though the USA and UK are regularly two of the largest recorded music markets in the world [3], Sweden, Switzerland, and Germany are the first that recognize popular patterns and lead the trends in the weekly top charts in the streaming world.

From this analysis, we can infer that if a song becomes popular in Spain, it will most likely succeed in the other Spanish speaking countries too. On the other hand, if it succeeds in Sweden, Switzerland or Norway, it might dominate the whole Western music scene. These findings can help greatly in predicting top charts of the follower countries, but also give better recommendations to users from those regions based on the most popular music in the leader countries.

6 Conclusion

The examination of the music streaming networks revealed clear evidence that communities are formed under the influence of language and geographic location. In addition, we presented an approach for detecting countries that lead the music trends on Spotify. The results showed that global leaders in the music industry are not necessarily trendsetters in the streaming world. Furthermore, it was revealed that both similarity and leadership between countries in music streaming differ across genres.

The results from this research could be leveraged from music streaming platforms, such as Spotify, to build recommendation engines based not only on content and/or collaborative-based filtering, but will also take into account specific demographic characteristics and music genre leaders and followers. In the future, we intend to perform this analysis using song-specific features. This approach will uncover new genres which are not merely labels, but a combination of certain feature values. Music has never been more available, thus the future awaits with many new challenges in the fields of music analysis and music recommendation.

References

1. Bastian, M., Heymann, S., Jacomy, M.: Gephi: an open source software for exploring and manipulating networks. In: Third international AAAI conference on weblogs and social media (2009)
2. Blondel, V.D., Guillaume, J.L., Lambiotte, R., Lefebvre, E.: Fast unfolding of communities in large networks. Journal of statistical mechanics: theory and experiment **2008**(10), P10008 (2008)
3. Defferrard, M., Benzi, K., Vandergheynst, P., Bresson, X.: Fma: A dataset for music analysis. ArXiv **abs/1612.01840** (2016)
4. Ferreira, F., Waldfogel, J.: Pop internationalism: has half a century of world music trade displaced local culture? The Economic Journal **123**(569), 634–664 (2013)

[3] `https://en.wikipedia.org/wiki/List_of_largest_recorded_music_markets`

5. Lao, J., Nguyen, K.H.: One-hit wonder or superstardom? the role of technology format on billboard's hot 100 performance (2016)
6. Lee, C., Cunningham, P.: The geographic flow of music. In: 2012 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining. pp. 691–695. IEEE (2012)
7. Menezes, R., Gunaratna, C., Patel, M.: Network sciences in music recommendation a case study with brazilian music (2012)
8. Nagy, M., Akos, Z., Biro, D., Vicsek, T.: Hierarchical group dynamics in pigeon flocks. Nature **464**(7290), 890 (2010)
9. Page, L., Brin, S., Motwani, R., Winograd, T.: The pagerank citation ranking: Bringing order to the web. Tech. rep., Stanford InfoLab (1999)
10. Salganik, M.J., Dodds, P.S., Watts, D.J.: Experimental study of inequality and unpredictability in an artificial cultural market. science **311**(5762), 854–856 (2006)