

# The Geographic Flow Of Music On Spotify

Miroslav Mirchev, Lidija Jovanovska


*Proceedings of the 16th International Conference for Informatics and Information Technology (CIIT 2019), Bistra, Mavrovo, Macedonia, 2019 – presented at the conference as a student paper, won 3rd prize*

## Cite this paper

Downloaded from [Academia.edu](#) 

[Get the citation in MLA, APA, or Chicago styles](#)

## Related papers

[Download a PDF Pack](#) of the best related papers 



[What Makes Popular Culture Popular?: Product Features and Optimal Differentiation in Music](#)

Noah Askin, Michael Mauskapf

[A COMPREHENSIVE STUDY OF THE POLISH MUSIC MARKET](#)

Katarzyna M. Wyrzykowska

[Artist Embeddedness and the Production of Novelty in Music](#)

Noah Askin, Michael Mauskapf

# The Geographic Flow Of Music On Spotify

Lidija Jovanovska

Faculty of Computer Science  
and Engineering

Skopje, Republic of North Macedonia  
lidija.jovanovska@students.finki.ukim.mk

Igor Mishkovski

Faculty of Computer Science  
and Engineering

Skopje, Republic of North Macedonia  
igor.mishkovski@finki.ukim.mk

Miroslav Mirchev

Faculty of Computer Science  
and Engineering

Skopje, Republic of North Macedonia  
miroslav.mirchev@finki.ukim.mk

**Abstract**—As Daniel J. Levitin interestingly noted, No known human culture now or anytime in the recorded past lacked music. Therefore, the impetus behind this research paper is to model the interactions between countries in order to reveal music listening trends at a macro level. Subsequently, the framework for performing this analysis consists of techniques used in the multidisciplinary field known as Network Science. Throughout the past decade, the world has witnessed a gradual shift in the way music is listened to. In that respect, Spotify, an online music streaming service, has been the imperative giant with a user base of around 191 million. With the help of Spotify’s Application Programming Interface (API), a dataset was compiled, which contains the Weekly Top 40 streamed songs across 50 countries, in the year 2017. Through research, the team explored whether, and to which extent, do language, nationality and geographic distance influence the way global communities are formed. Furthermore, the project aimed to prove that there is a clear direction of leadership flow in the network. Until now, the acquired information supports the hypotheses that some countries do indeed follow the trends beset by others and that language and nationality play an essential role in the development of communities.

**Index Terms**—music, network science, clustering, leadership

## I. INTRODUCTION

Even though music is deeply rooted into mankind’s history, people began recording it in the mid-1850’s. After years of improving and innovating on recording technologies, mankind reached the Internet era, which increased the ease of access per user, as well as enrich the music world with a wide variety of artists and genres. As the efficiency and availability of physical music recordings waned in recent years, a substantial part of the population decided to cross over to music streaming services. While physical and download revenue continue dropping worldwide, digital and streaming services are experiencing the exact opposite. Streaming revenue increased by 230% between 2013 and 2017, marking it as a period of consistent growth.

Spotify is undoubtedly the most popular platform in the music streaming business, harboring a user-base of around 191 million. The platform also provides a powerful Application Programming Interface (API), through which Music Information Retrieval (MIR) researchers can gain access to audio features, artist information, as well as daily and weekly global streaming charts. Due to Spotify’s dominant position in the business, the data analyzed in this project was acquired

through their API because it provides current relevant information regarding each country’s streaming preferences.

Lao and Nguyen analyzed data from the ‘Billboard Hot 100’ charts spanning from 1958 to 2015. They found that the switch to digital technology significantly lowers the cost of publishing a single hit, enabling already popular artists to increase their fame and consequently homogenize the top charts. As Ferreira and Walldogel found in their 2010 research on the global music trade, from 2001 to 2007, 31 artists have appeared simultaneously on at least 18 countries’ charts in at least one year [1]. However, digital songs are more likely to fall off the chart in the first week compared to CD songs, which indicates a highly volatile nature, resulting in heterogeneous charts on a regular level [2].

Our analysis of the geographic distribution of musical preferences is structured as follows: We begin by describing the data, a world-wide log of streaming habits recorded by Spotify, as well as various preprocessing steps in Section III. In the following step, in Section IV, we describe how we constructed the network and investigate whether and to what extent communities are formed based on language and geographic distance.

In Section V we describe how we adapted and adjusted a methodology previously used to find leadership in pigeon flocks to detect leader-follower relationships between countries. The leaders and followers methodology involves examining every dyad between the countries in the dataset and testing whether the time-lagged correlation is significantly larger in one direction than the other.

## II. RELATED WORK

The field of Network Science provides a framework for modeling interactions between entities so as to reveal properties at a macro level, which may not be noticeable or visible at the individual level. Techniques from this area of study have been successfully implemented to many other fields including Music [3].

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 [4]. This was further adapted by Lee and Cunningham in their research about the global flow of music on Last.fm [5], which served as the incentive for this paper.

Shafiq, Ilyas, Liu and Radha developed a model for identifying specific types of leaders, followers and neutrals. The model was applied on data from Facebook and was ultimately able to capture the characteristic differences of the user categories [6].

### III. DATA

#### A. Preprocessing

Spotify keeps record of the daily Top 200 Charts for every available country on a dedicated site. The data can be scraped systematically by downloading the .CSV files separately for each country. The scope of this analysis will be the entire 2017 year from January 1<sup>st</sup>, 2017 to January 1<sup>st</sup>, 2018, inclusive.

Because not all Spotify users are consistently daily active, a single day's chart can be thought of as a sample of listening preferences among users. In countries that have relatively few users, the variance associated with this sample becomes large, indicating high volatility. This noise can be reduced by aggregating a matrix associated with seven consecutive days together. By doing that, it effectively increases the sample size for each entry in the country-song matrix.

For some countries, there was not sufficient data to compile a Weekly Top 200 list for every week in the entire year. Since record charts have traditionally consisted of a total of 40 songs, it was decided that a Weekly Top 40 Chart will represent a country's weekly streaming preferences accurately and avoid making a significant reduction to the country list.

#### B. Missing Data

Despite downsizing the dataset, there was no sufficient data to compile the Weekly Top 40 lists for several of the countries that newly adopted the service, including Lithuania, Luxembourg and Estonia. Consequently, those countries were removed from the dataset, resulting in a final list consisting of 50 countries. By doing this, the size of the dataset reduced significantly, while the number of unique songs dropped from 21,747 to 1551, indicating that the Top 200 Charts are significantly more heterogeneous than the Top 40 Charts.

#### C. Creating Streaming Matrices

In order to create a suitable representation for a country's streaming history, we aggregate the data in so called 'streaming matrices'. In that manner, we have a matrix for each week. In this matrix every country is a row vector with 1551 elements, and each column represents a song. Each song in the vector defines a dimension in Euclidean space and the frequency of each song corresponds to the value in that dimension. Since not all countries have the same set of songs appearing on their Top 40 chart, there is a large number of zero-valued elements, resulting in sparse streaming matrices. Therefore, 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  $j$  in that particular week.

### IV. CLUSTERING: GEOGRAPHICAL CLUSTERS ARE STRONG

#### A. Creating the network

With the purpose of comparing the streaming vectors of each pair of countries, we must choose a similarity measure.

Not all similarity measures yield the same results in a certain scenario. Therefore, it is essential to choose the metric that will describe the data properly. In this project, we considered the use of two measures: cosine similarity and jaccard similarity.

By using jaccard similarity, we risk losing information regarding the songs' streaming frequencies that may prove to be useful when deciding how to generate the network. During testing, cosine similarity yielded more plausible results, as expected, and was subsequently used when measuring the similarities of each pair of countries' streaming preferences.

**Determining the threshold.** In our graph, each node represents a country, while each link between a pair of nodes indicates that those countries have a similarity index above a certain threshold. The determination of the threshold in this work is not validated by human subjects, nor estimated by a predictive model. The value of 0.3 for the threshold was obtained as a result of extensive testing.

Since 2014, the average number of artists each Spotify user streams per week has increased by 37 percent. So far, in 2017, it rose from just under 30 to about 41 different artists per week. This fact might lead us to believe that the rise of streaming services has increased heterogeneity across musical charts. To test this claim we attempt to find whether clusters are formed based on language and geographic distance.

To construct the dendrogram shown in figure 1, we performed average linkage clustering (an agglomerative clustering algorithm) on the adjacency matrix  $A$  of the countries, a square matrix where each entry  $A(i, j)$  is the cosine similarity between country  $i$  and country  $j$ . Instead of constructing the dendrogram based on just a single streaming matrix, we summed together the similarity matrices spanning from January, 2017 to December, 2017.

#### B. Discussion

Starting from the lowest level structure of the tree, we can already see clusters of pairs of countries which are geographically close to each other. Such examples include: The United Kingdom and Ireland, New Zealand and Australia, Germany and Austria, Slovakia and the Czech Republic, Bolivia and Ecuador, Chile and Peru etc. There are some exceptions to this claim, notably Latvia, which seems to be more similar to New Zealand and Australia, than to other European countries. For example, it is strange to see The United States being closer to Iceland, rather than to The United Kingdom, despite the fact that Iceland is the closest European country to The United States.

At an intermediate level, we can observe that most clusters are formed based upon language. There are two clusters

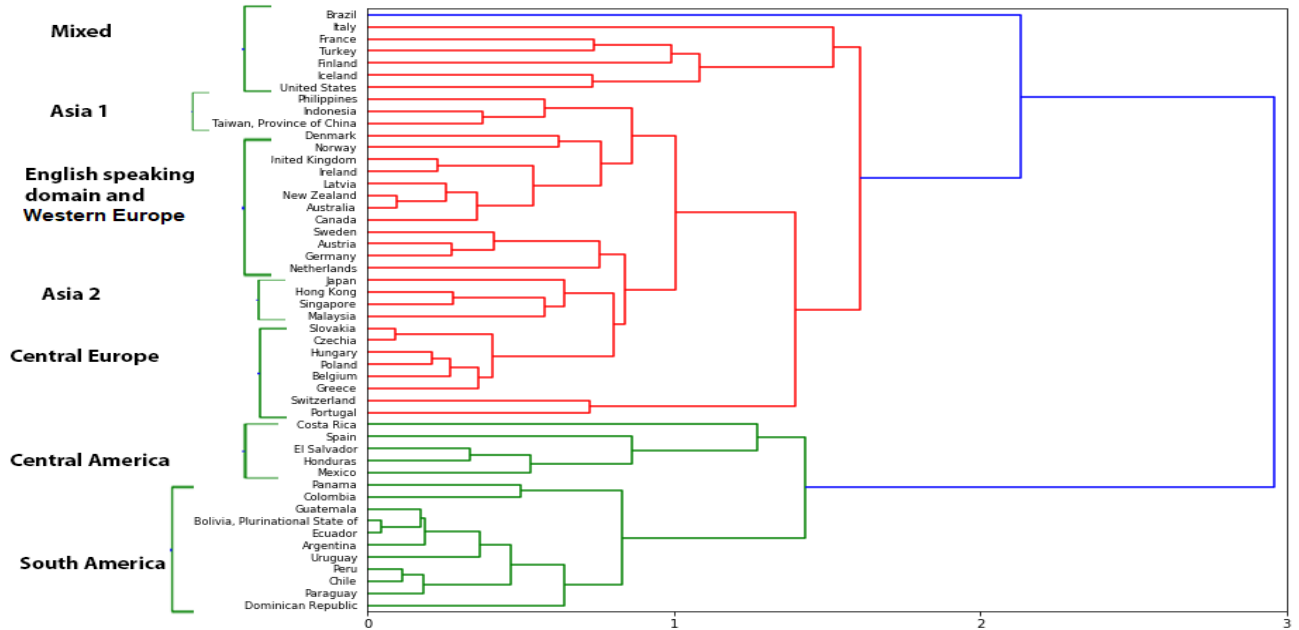


Fig. 1. Hierarchical clustering based on average linkage clustering of cosine similarities of countries in the normalized streaming matrices.

of Spanish speaking countries; one consisting of Spain and countries from Central America, and another, consisting of countries from South America. An English dominated cluster emerges, with Latvia being the odd one out. The Eastern Asian Countries form two clusters which despite their geographical adjacency, are not close at all regarding their similarity scores. The rest of the clusters include Slavic speaking countries, along with Greece, Hungary and Belgium, and three clusters consisting of Western and Northern European countries. It is surprising that Switzerland and Portugal form a cluster even though there is no overlap in the sets of nationalities that inhabit those countries, nor is there a common language. The multinational character of Switzerland would be expected to be a key factor in high similarities between Germany, France and Italy. The strangest cluster at this level is definitely the one constituted of countries such as: Iceland, United States, Turkey, Finland, France and Italy.

At the next level of the hierarchy, there are two notable clusters. Here, language seems to play a key role in the formation of these clusters. The first cluster consists exclusively of Spanish speaking countries, while the other includes every country left in the dataset, except Brazil. We would expect Brazil to be relatively closer to the first cluster, due to the geographical proximity and the fact that Portuguese and Spanish are similar languages. However, to our surprise, Brazil's community joins the second cluster on a relatively higher level. We speculate that Brazil does not belong to a

particular community because it's the largest country in South America and hence it might have its own distinct idiosyncratic preferences.

## V. LEADERS AND FOLLOWERS

To detect leader-follower pairs, the methodology of Lee and Cunningham was implemented, which is based on finding lagged correlations, and was also previously applied to finding leadership in global music listening preferences. When examining the relationship between a pair of countries, we are interested whether there is a directed link from one country to the other, or whether it is of neutral nature, where neither leads the other.

We begin by calculating the velocities for each country in the dataset. A velocity  $v_{country}(t, t+1)$  represents the change that takes place in the listening habits of a country from one week  $t$  into the next  $t+1$ . That leaves us with a sequence of velocities for each country.

To measure whether a country  $i$  follows a certain country  $j$ , we measure the cosine similarity of each of the country  $i$ 's velocities with the velocities of the country  $j$  from one week earlier. The average of these lagged similarities are referred to as the correlation of the first country's velocities with the second country's lagged velocities, where the lag size is one week. We call this measure  $C$ .

### A. Deciding Which Edges To Accept

It could be the case for a dyad  $i, j$  that after testing whether a correlation is strong enough to be accepted,  $i$  appears to follow  $j$  and  $j$  appears to follow  $i$ . While it would be easier to choose the direction with the higher correlation, that option could mean that neither country is leading another, and instead, they are moving together.

To make sure there is a clear direction to the leader-follower relationship, we perform a t-test to make sure that the two correlations (which are means of similarities) are not equal; here we use a two-sided, paired t-test. Since our sample size is relatively small, we perform a Levene test to check whether the variances of the correlations are equal. If the test is positive, i.e. the null hypothesis is accepted, we proceed with the independent t-test. Otherwise, we use the t-test for related samples. If one correlation is larger, then we accept the leader-follower pair associated with that correlation as a directed edge; otherwise, it is concluded that no leader-follower relationship exists.

## VI. RESULTS

After adapting the methodology to the subject at hand, we find all leader-follower relationships in the network and then assign edges weighted by the lagged correlation. The graph which can be seen in Figure 2 is a DAG (directed acyclic graph). The size of the nodes indicate their PageRank and the color represents the community to which they belong. In the past, it has been argued that a system with a strong leadership hierarchy ought to be nearly acyclic, so the lack of cycles in the output network is a clear validation of that theory [7].

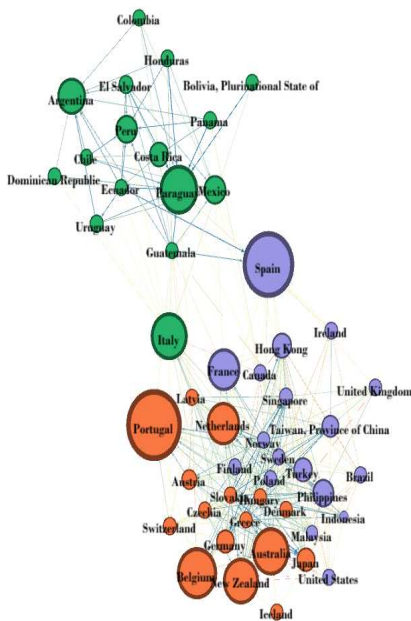


Fig. 2. The Geographic Flow of Music on Spotify

There are many centrality measures that could be used as criteria for deciding which countries are musical trendsetters and which are laggards. PageRank was used because it is an algorithm designed to rank importance of nodes on weighted, directed networks on which a dynamic process takes place [8].

As we can see in Table I, Spain and Portugal are the most important nodes in the network due to the fact that they both serve as bridges between the South American and the Eurasian communities. This indicates that language and geography play a key role in streaming preferences, since both European countries use Latin languages. Both communities have no outgoing links and yet they are followed by some of the more active countries including The United Kingdom, Sweden, Norway etc. It is strange to see Australia and New Zealand's prominent rank, while we would expect The United States, United Kingdom and Japan to fare much better.

TABLE I  
STATISTICS FOR THE MOST IMPORTANT NODES IN THE NETWORK

Country	PageRank
Portugal	0.068441
Spain	0.062991
Belgium	0.047724
Paraguay	0.045235
Australia	0.043667

## VII. CONCLUSION AND FUTURE WORK

This exploratory data analysis aimed to find out whether countries form streaming communities based on language and geographic distance. There is evidence to support this claim, although there are a few exceptions. Furthermore, two more aspects give credibility to the results: that each leader-follower relationship underwent a t-test, and that when all of the leader-follower relationships were put together into a graph, they formed a DAG, indicating a direction of flow in a strict sense.

Future plans include obtaining a larger dataset and implementing the leaders and followers methodology on sets of songs belonging to various genres. By doing so, we can shed light on trends, hidden, due to the multi-dimensional aspect that genre brings to the data.

## REFERENCES

- [1] F. Ferreira and J. Waldfogel, "Pop internationalism: has half a century of world music trade displaced local culture?," *The Economic Journal*, vol. 123, no. 569, pp. 634–664, 2013.
- [2] J. Lao and K. H. Nguyen, "One-hit wonder or superstardom? the role of technology format on billboards hot 100 performance," 2016.
- [3] C. A. Perrone and C. Dunn, "Brazilian popular music and globalization," *Journal of Popular Music Studies*, vol. 14, no. 2, pp. 163–165, 2002.
- [4] M. Nagy, Z. Akos, D. Biro, and T. Vicsek, "Hierarchical group dynamics in pigeon flocks," *Nature*, vol. 464, no. 7290, p. 890, 2010.
- [5] C. Lee and P. Cunningham, "The geographic flow of music," in *2012 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, pp. 691–695, IEEE, 2012.
- [6] M. Z. Shafiq, M. U. Ilyas, A. X. Liu, and H. Radha, "Identifying leaders and followers in online social networks," *IEEE Journal on Selected Areas in Communications*, vol. 31, no. 9, pp. 618–628, 2013.
- [7] E. Mones, L. Vicsek, and T. Vicsek, "Hierarchy measure for complex networks," *PloS one*, vol. 7, no. 3, p. e33799, 2012.
- [8] L. Page, S. Brin, R. Motwani, and T. Winograd, "The pagerank citation ranking: Bringing order to the web," tech. rep., Stanford InfoLab, 1999.