Analysis of geographical and sentiment patterns in a network of songs

Matija Sipek[1]**, Alejandra Navarro Castillo**[1]**, and Martin Karlik**[1]

[1]DTU Technical University of Denmark

This manuscript was compiled on December 10, 2022

In the modern world, music is an omnipresent occurrence, and our days start and end with it. It is therefore interesting to learn how music shapes people, in particular communities of people who listen to a similar type of music. Using network science tools, we create an undirected network of 1766 songs that are popular in any country in the world, and we link similar songs together, ending up with 3569 edges. Extracting and analysing the communities of our network provides an interesting insight into the popularity of various genres in different regions of the world. We then calculate the sentiment values for the lyrics of the songs in our network. Finally, we statistically evaluate, whether there is a correlation between the community's average sentiment and the community's average GDP per capita, or the average yearly temperature. With reasonable doubt, we discover that people living in countries in colder climates tend to listen to less joyful music, compared to the rest of the world.

network science | graph theory | social graphs | music sentiment | music communities

Introduction

How humans interact with each other and join a community is always an interesting question, not only for scientific purposes but also for every person's life. These interactions depend on a person's personality and habits and on the culture and traditions of a society. One aspect that might be an indicator of how these communities are modulated is music. Using graph and network theory tools, we have been able to analyze a graph of songs (with attributes such as genre tags, sentiment, and a list of countries where the song is popular) and identify a number of communities. Also, we analyze how the different types of music are divided and get insightful information about various characteristics of each music community. Our final goal is to analyze whether the communities that we obtain from the graph have a correlation with country information such as gross domestic product (GDP) and the average temperature throughout the year. One of our research questions is; could it be that countries with warmer climates tend to listen to happier music, or perhaps the other way around?

Data Preparation

We used several datasets to reach the goal: MillionSongsDataset (MSD) as the core of our system, this is a large online database of multiple music related datasets, and the basis for our graph. Next, the musiXmatch lyrics dataset for sentiment analysis and extracted GDP and temperature data for each country using Wikipedia.

Songs dataset. The core dataset from MSD we used is Last.fm, and the data in this subset is in the format artist, timestamp, similars, tags, trackID, and title with an example in Table 1. We are using the *similars* attribute to create edges between artists, and for each artist we are adding tags. Also, there are song-level tags and similar songs. The algorithm used for similarity measurement is however not public and we could not find any information on it.

Name:format	Example data
artist:string	Jos Merc
timestamp:time	2011-08-16 01:34:38.887856
similars:string/int	["TRNIEVD128F147645F", 1], ["TRZIWPD12903CDE96C", 0.66]
tags: list of strings	[Flamenco,world,cante flamenco,jose merce]
trackID:string	TRVABRY128F1476445
title:string	Campesino y minero (Tarantos)

Table 1. Dataset row example.

Choosing a subset. The dataset of all songs ever recorded is very large that it is practically impossible to consider it in its entirety to conduct our study; in practice, we are forced to choose a subset. Because of the computational limitations of our academic environment, we initially considered a random subset of 10,000 songs from the Last.fm dataset. Upon reflection on this approach and its arbitrariness, we aimed to instead choose a subset that could be motivated by certain criteria related to our study. Since we are interested in the songs' popularity in the different countries, we ended up choosing

Significance Statement

There are various elements that describe a human society. One of them may be the type of music they perform or listen to. Thus, in this paper we focus on how different types of music can be related to different societies (or communities) and analyze what aspects of this communities are involved in the choice of their music. For this purpose, we study whether the GDP, weather or location of countries have some relationship with the kind of music their inhabitants listen to. Analyzing this question gives us consciousness about the reasons why each type of music or artist is popular in each region, and it is also useful for artists to know where they could find their best audience based on the type of music they play.

All authors contributed equally to this work. Martin K. was mostly responsible for lyric and country data preparation, along with sentiment analysis, Matija S. for the song dataset preparation, network creation and community detection and Alejandra N. C. for network analysis, and statistical evaluation. But overall, everybody took big interest in every part of the project.

Table 2. An example illustration of the lyrics data from the database.

track_id	word	count
1. TRAAAAV128F421A322	love	11
2. TRAAAAV128F421A322	and	5
3. TRAAAAV128F421A322	heart	3

only the songs that are among the top 100 most popular songs in any country. This way, we ended up with a final subset of 1766 songs.

Lyrics dataset. The MSD platform also provides a dataset of lyrics as a sqlite3 database, with the data given as illustrated in Table 2. This dataset, however, does not perfectly overlap with the dataset of songs from Last.fm; only approximately 25% of songs from the original subset were found in the database of lyrics. After taking a subset based on the popularity of the songs in all the countries, 51% of the songs were found. Although the subset split is still largely arbitrary, having a subset of popular songs did help increase the amount of our lyrics data.

The track_id format is equivalent to the song ID format in the Last.fm dataset, therefore, we used it to query the database to get the lyrics for the songs in our subset. The lyric data, provided as a term-frequency dictionary (a stemmed word and its count), was eventually used to calculate the sentiment value for each song.

Countries dataset. To construct a dataset of countries with all the relevant information, we first used a plain list of countries as a text file (1) There is unfortunately no true list of countries, as they depend on who is creating the list, therefore, ours is also just an approximation. We then extracted the information on GDP per capita and average yearly temperature for each country from the internet (2), (3). We stored the final dataset as a text file, as illustrated in Table 3.

Table 3. An example from the countries dataset.

country name	GDP per capita	average yearly temperature (ºC)
1. Netherlands	1,018,007.06	9.25
2. New Zealand	249,991.51	10.55
3. Nicaragua	14,013.02	24.90

Network

As discussed in Section *Choosing a subset*, we chose a subset based on the song popularity in each country, which resulted in 1766 songs. In order to create the network, we decided to use the songs as nodes. To create the edges, we chose to use the neighbours provided by the Last.fm dataset, calculated using a similarity measure, as discussed in Section *Songs dataset.* Our final network is visualized in Figure 1, using the forceatlast algorithm.

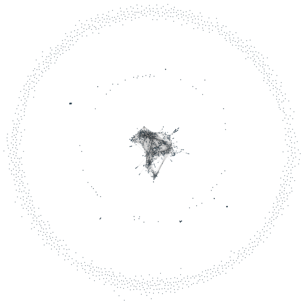


Fig. 1. Plot of the whole network.

In addition, for each node in the network, there is a set of attributes which are: 'artist', 'title', 'timestamp', 'tags', 'sentiment', 'countries', 'gdp', 'temperature' and 'neighbors'. The attributes 'artist', 'title', 'timestamp' and 'tags' were simply taken from the Last.fm dataset. The attribute 'countries' was constructed as the list of countries, where the song is among the top 100 most popular songs. The attributes 'gdp' and 'temperature' were calculated from our country dataset, taking the average of these values, when considering the entire list of node's countries. Lastly, the 'sentiment' attribute of each node was calculated using sentiment analysis.

Sentiment Analysis. We used the lyrics database to get the lyrics data for each node in our network. For a song that was found in this database, we computed the overall sentiment by comparing each word in the found lyrics against a list of sentimentally annotated words of English language, created by Dodds et al. (4). We disregarded the stopwords, in order to focus attention on the more unique and characteristic words in the lyrics. We collected all the words in the lyrics, for which we found a match in the list of sentimentally annotated words, accessed the sentiment value for the given word (1 to 10), and finally computed the average sentiment value for every found word in the lyrics of the given song.

One problem we needed to address was that the lyrics database contains the words in a stemmed form; however, the exact stemming technique was not provided. To properly compare the stemmed lyrics data against the unstemmed sentimentally annotated list of words in cases where the words should match (e.g., "unique" and "uniq"), we stemmed the list of sentimentally annotated words.

Network Analysis. After having created the network, we obtained the following results: The number of nodes in the network is 1766, and the number of edges is 3569. The maximum degree is 51, and the minimum is 1.

Studying the degree distribution of the network (Fig. 2), we can see that on a log-log scale the data points form an approximate straight line, so the degree distribution can be approximated with a power-law distribution with degree exponent $\gamma = -1.571$ (Fig. 3). Hence, the network is a *scale-free network.* The degree distribution follows the equation in 1.

$$p_k = Ck^{-\gamma} \qquad [1]$$

Communities. In a network, communities are locally densely linked subgraphs. For community detection, we are using the *community package*, and also to measure modularity, we get a modularity result of 0.5367 indicating a community structure. So, we implemented the Louvain algorithm to find a good

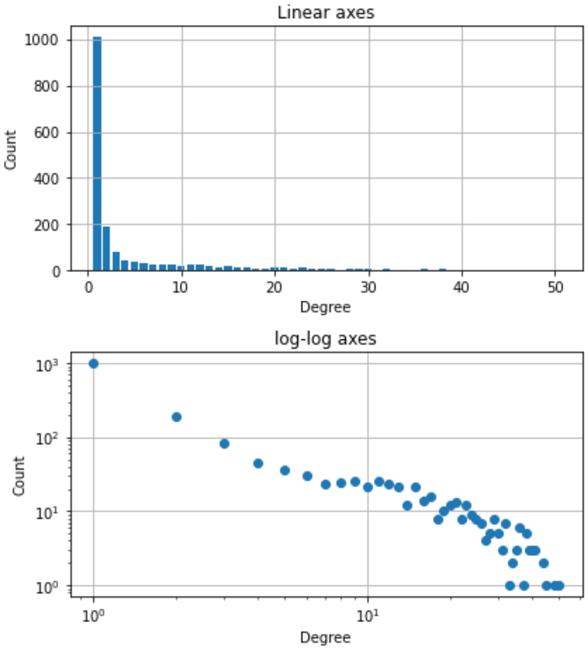


Fig. 2. Degree distribution of the whole network in both linear and loglog axis.

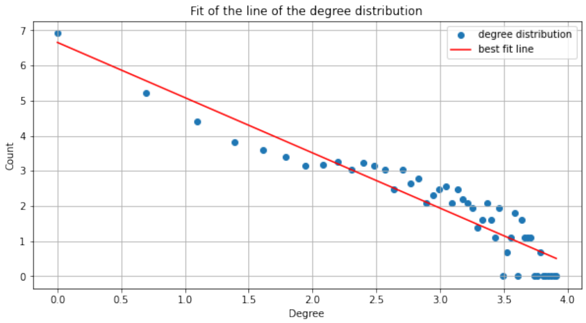


Fig. 3. The degree distribution is shown on double logarithmic axis (log-log plot), in which a power-law follows a straight line.

split. The results are as follows; the number of communities in the network is 66; the largest three contain (133, 104, and 78) nodes and the smallest are 2 (we removed isolated nodes). As can be seen from the graph, the largest communities are all part of the giant component in the center, with a number of smaller communities surrounding them.

Natural Language Processing (TF-IDF). TF-IDF stands for term-frequency (TF) inverse-document frequency (IDF), and it is a computational technique for measuring the relevance of a word to a document in a collection of documents. The first part, TF measures frequency of a word in a document, and IDF how rare a word is withing a set of documents. Due to the format of our dataset, we are able to leverage this technique by measuring song tag levels. Since, an artist has multiple songs, and each song has multiple tags, we are able to calculate the relevance of a tag within a corpus of artists' songs. And, similarly IDF compared with all songs within the graph. Our goal was to find some correlation in data withing communities, so TF-IDF was calculated for each community, and word clouds and frequency graphs to visually inspect the differences and trends.

Results

Visual interpretation of communities. The detailed visualization of the 16 largest communities is shown in Appendix A, in Figures 5 and 6.

Visual inspection suggest that generally, the tags in each community seem to be centered around a certain genre or a theme. However, s as for the country word clouds, the popularity of most song-communities seems to spread around multiple continents. We discuss a few instances, where we can speculate on some patterns in the behaviour of the country data, and its correlation with the community's tag data.

In Figure 6, the community #11 demonstrates that music tagged as "rockargentino" or "latinrock" is, as expected, dominantly popular in Latin America. The community #12, containing tags such as "finnish", "suomi" or "suomirock" is most popular in Finland, but interestingly also in Thailand, Bhutan and Dominica. Community #16 shows that songs tagged as "christmas", "xmas", "christmasmusic" are especially popular in Europe, with Denmark being the top listener to Christmas music. This finding is interesting, and begs to more closely look at the Danish culture, perhaps the Danish concept of "hygge" (5) plays or role, or the fact that Christmas is celebrated earlier in Denmark than in most other European countries – but that is only speculating, and an interesting concept to study in future work.

The rest of the communities don't seem to suggest any clear geographical clustering, but we invite the readers to inspect the communities in Appendix A by themselves.

As for the correlation of the country data and the average sentiment of the songs, we cannot conclude much by individually inspecting each community – this is where we opt for a statistical evaluation.

Statistical evaluations. To further inspect other aspects of the network, we decided to investigate the following hypotheses.
H_1 : songs that are popular in countries with a "high" temperature have a higher average value for sentiment than a similarly sized set of randomly selected songs.
H_2 : songs that are popular in countries with a "low" temperature have a lower average value for sentiment than a similarly sized set of randomly selected songs.
H_3 : songs that are popular in countries with a "high" GDP have a higher average value for sentiment than a similarly sized set of randomly selected songs.
H_4 : songs that are popular in countries with a "low" GDP have a lower average value for sentiment than a similarly sized set of randomly selected songs.

We have performed a permutation test for each hypothesis. The permutation test is a nonparametric form of a statistical inference test where we resample from the data without replacement. To do so, we get a sample R of the whole graph G of the same size as the subgraph that we want to test and we compare how the average sentiment of R is different than the average sentiment for the subgraph. We repeat the same process a considerable number of times (in our case 5000 iterations).

Table 4 summarizes the results of each of the statistical tests.

Table 4. p-values for each of the statistical tests.

H_1	H_2	H_3	H_4
0.383	0.039	0.864	0.395

Looking at the p-values of each of the test, we can conclude that for hypotheses H_1, H_3 and H_4 there is not enough evidence to accept them. However, in the case of hypothesis H_2, we see that the p-value is 0.039 which is smaller than the significance level of the test ($\alpha = 0.05$). That means that we can conclude that "songs that are popular in countries with a *low* average temperature have a lower average value for their sentiment compared to a similarly sized set of randomly selected songs". In figure 4 we can see the permutation test distribution for hypothesis 2.

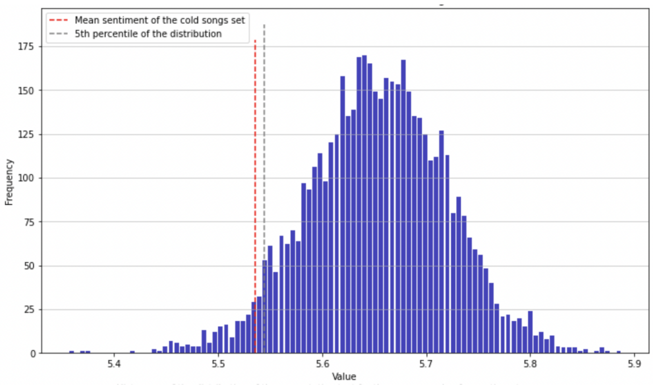


Fig. 4. Permutation test distribution for H_2. The red line represents the average sentiment for the songs that are popular in countries with a *low* average temperature. The grey line represents the 5th percentile of the distribution.

Methods

Packages and libraries. To achieve our goals, we used multiple packages and libraries. The core of our system is the network analysis package NetworkX. It is used for graph based structures, and key elements are nodes and edges. The next step was to visualize our graph. For that we used firstly NetworkX built-in function, but later we used ForceAtlas2, which is a graphical layout algorithm for force-directed graphs, as it is more advanced and provides better visualization. For natural language processing (NLP) we used Python package called Natural Language Toolkit (NLTK) used to preprocess and analyze text. It has a number of libraries for symbolic and statistical NLP such as; classification, tokenization, parsing, stemming etc. From the graph created by NetworkX, we found communities using the Louvain Community Detection Algorithm. This algorithm is a two step process: in the first part each node is defined as its own community, and in the second step it is moved to neighborhood communities and positive gain in modularity is measured. We have used Python programming language and Jupyter Notebook as a computing platform. For saving structural properties of a graph we have used GraphML, and for data analysis and manipulation we have used Pandas.

All the implementation and data analysis can be found in our Jupyter Notebook, available on Github (6).

Discussion

In reflection, we discovered several insights into the world of music, and how people around the world listen to it, but some of our methods should be called into question.

First, our technique to assess the sentiment of each song is suboptimal in a number of ways: by using only the lyrics to assess the sentiment, we disregard all the other relevant musical information, which is very much relevant to the overall sentiment. Another problem was the lack of data: not only did we not have access to all the songs' lyrics, but also even for the lyrics we could access, we could use only a subset of them to calculate the sentiment – those for which we found a match in the sentimentally annotated list of words. Also, our technique of stemming could be different from the one used for the lyrics database, therefore missing cases, where a match should occur (e.g. "uniqu" and "uniq") – however, we did not evaluate if this is the case, or how often we miss such matches.

Our choice of the subset – selecting the popular songs in each country – is still largely arbitrary, e.g. having a limit of one hundred popular songs. With this in mind, we cannot confidently speculate how well our results generalize to the entire dataset of songs, as we chose only a rather arbitrary window. Instead, a different sampling technique, such as the Snowball Sampling Completion (7) could be considered.

For future work, it could be interesting to see how the communities form, when taking time information into account: what kind of music was popular throughout history, and do the countries evolve to obtain different taste in music? Does the sentiment of music correlate with historical events, such as wars? Or on a smaller time scale, how does the countries' taste in music change throughout the year, depending on the season? These are all interesting questions that the scope of our project did not cover. But we hope that our project sparked an interest in this topic, and can be expanded upon in the future.

References

1. List of countries: https://gist.github.com/kalinchernev/486393efcca01623b18d (2022).
2. Gdp per capita for countries: https://data.worldbank.org/indicator/NY.GDP.MKTP.CD (2022).
3. Average yearly temperature for countries: https://en.wikipedia.org/wiki/List_of_countries_by_average_yearly_temperature (2022).
4. PS Dodds, KD Harris, IM Kloumann, CA Bliss, CM Danforth, Temporal Patterns of Happiness and Information in a Global Social Network: Hedonometrics and Twitter. *PLOS ONE* **6**, e26752 (2011) Publisher: Public Library of Science.
5. M Wiking, *The little book of hygge: Danish secrets to happy living.* (William Morrow, an imprint of HarperCollins Publishers, New York, NY), (2017) OCLC: ocn951936216.
6. Jupyter notebook (2022) https://github.com/martinkarlik/Geographical-Patterns-in-Music.
7. IM Dragan, A Isaic-Maniu, Snowball Sampling Completion. *J. Stud. Soc. Sci.* **5** (2013).

A. Analysis of the sixteen largest communities

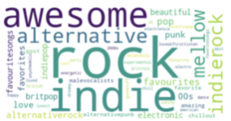
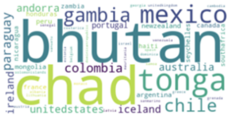
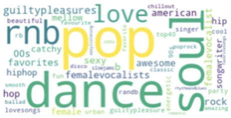
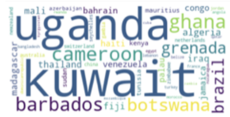
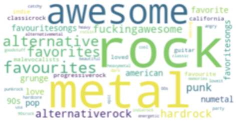
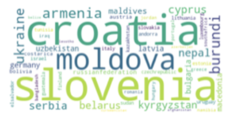
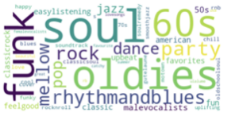
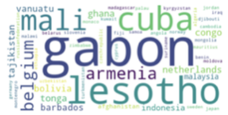
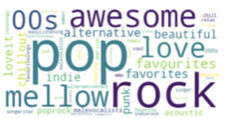
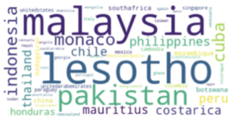
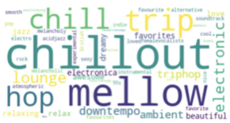
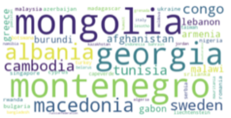
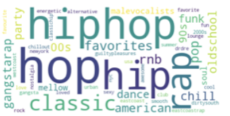
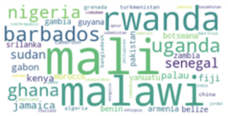
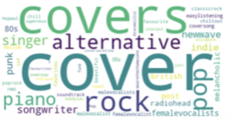
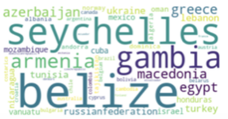
Community No.	Tag word cloud	Country word cloud	Sentiment	GDP per capita	Temperature
#1			5.70	225,541$	17.86
#2			5.81	256,005$	25.13
#3			5.57	56,419$	8.14
#4			5.85	173,681$	20.82
#5			5.73	233,470$	21.56
#6			5.74	80,562$	12.40
#7			5.59	68,213$	25.07
#8			5.83	337,679$	14.97

Fig. 5. Communities #1 to #8. The communities are ordered from the largest (the most amount of nodes) to the smallest. The tag- and country- word clouds are constructed using TF_IDF; the more frequent terms are visualized bigger. Sentiment is calculated as the average value of the sentiment of each song in the community. GDP per capita and (average yearly) Temperature are calculated as the average value for each country in the community.

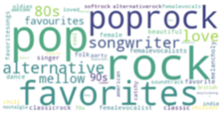
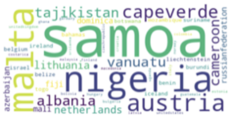
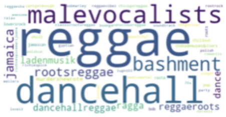
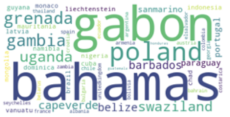
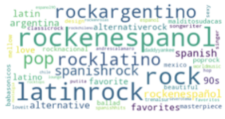
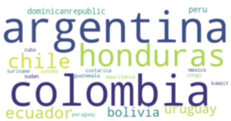
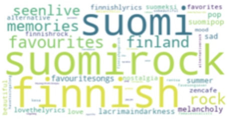
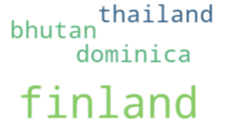
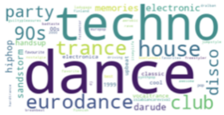
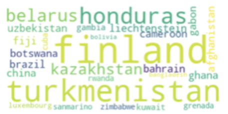
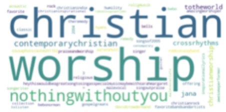
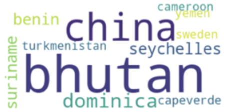
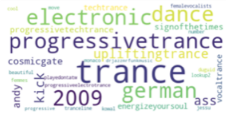
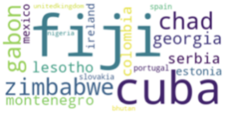
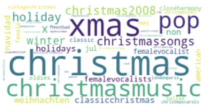
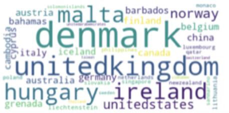
Community No.	Tag word cloud	Country word cloud	Sentiment	GDP per capita	Temperature
#9	 pop rock favorites poprock 80s favourites songwriter love alternative mellow 90s	 samoa malta nigeria austria tajikistan capeverde cameroon vanuatu lithuania albania netherlands azerbaijan mali russianfederation	5.70	55,679$	23.55
#10	 reggae dancehall malevocalists bashment rootsreggae jamaica ladenmusik dancehallreggae ragga reggaeroots dance	 gabon bahamas poland grenada gambia uganda barbados paraguay capeverde belize swaziland liechtenstein sanmarino guyana monaco portugal	5.68	100,723$	22.27
#11	 rockargentino rockenespanol rocklatino latinrock spanishrock rock pop latin argentina alternativerock spanish alternative favorites masterpiece rockenespañol	 argentina colombia honduras chile ecuador bolivia uruguay dominicanrepublic peru	4.99	178,210$	20.16
#12	 suomi suomirock finnish seenlive memories favourites finland rock favouritesongs finnishlyrics summer zencafe melancholy lacrimaindarkness		5.07	201,999$	14.40
#13	 techno dance trance house eurodance club party 90s disco hiphop sandstorm memories electronic darude	 belarus honduras finland turkmenistan uzbekistan fiji botswana brazil china kazakhstan bahrain ghana afghanistan gabon liechtenstein cameroon gambia sanmarino zimbabwe kuwait grenada luxembourg	5.62	1,978,896$	18.70
#14	 christian worship nothingwithoutyou contemporarychristian crossrhythms totheworld jana	 china bhutan benin suriname turkmenistan seychelles dominica capeverde cameroon yemen sweden	5.81	1,849,786$	18.29
#15	 electronic dance progressivetrance upliftingtrance trance german 2009 kick andy cosmicgate signofthetimes ass energizeyoursoul vocaltrance	 fiji cuba gabon chad georgia serbia estonia zimbabwe montenegro lesotho mexico ireland colombia portugal slovakia spain bhutan	?	308,360$	15.93
#16	 christmas xmas christmasmusic pop christmassongs holiday winter holidays christmas2008 femalevocalists femalevocalist classicchristmas weihnachten	 denmark unitedkingdom malta hungary ireland norway belgium china austria bahamas barbados monaco cambodia italy iceland australia germany finland canada grenada liechtenstein unitedstates	6.18	2,979,148$	11.65

Fig. 6. Communities #9 to #16. The format of the displayed information is detailed in Figure 5