Community Detection and Friendship Recommendation in online Music Streaming Services

Chun-Han Chang
Julien Crabié
Paul-Henri Castets
Oleg Struyanskiy

ABSTRACT

Deezer is the one of the largest music streaming services in the world and offers a variety of music content from labels including Sony Music, Universal Music Group, and Warner Music Group on various devices online or offline. As its competitors, Deezer also provides a recommendation system to improve the musical experience of its users, making them discover or rediscover songs related to those they are already listening to.

The goal of this project was to further improve the users' experience by improving the recommendation system itself. This project aimed at finding experts among users and linking them to other users, in order to suggest them more personal and tailored recommendation, instead of algorithm-based recommendations.

To achieve this ambitious goal, several graph-mining methods were used. In particular, the Louvain method and graph sparsification were applied for community detection. Degree centrality and maximum degree were also used to detect the experts inside the communities.
The results obtained suggest that each community listens to multiple musical genres. Most communities listen to 'universal genres' (e.g. pop music) and are differentiated by the number of listeners of less 'universal genres' (e.g. reggae). The degree distribution of the dataset was very skewed and nodes with larger degrees were rare, which facilitated the experts detection. The detected experts, as they belong to their individual community are considered relevant advisor as they have similar tastes as other users belonging to the same community and should help improve the global user experience.

1 INTRODUCTION

1.1 Motivation

In spite of existing research on Music Streaming Services (MSS), the research that was found was mostly dealing with recommendation systems of those platforms. This project will aim at partly completing this existing work.

From a user perspective, recommendation systems are an essential tool, but it still lacks precision in its suggestions and the results it provides can be disappointing. As algorithms, the recommendation systems also lack the human side that a person-to-person recommendation would provide.

Human recommendation are not necessarily more accurate and don't necessarily lead to better results, but the level of trust is higher between two individuals and one will tend to follow another person's recommendation more often than an algorithm's recommendation. Further friendships could also emerge between 2 passionate music listeners, thanks to community belonging feelings and personality appreciations.

1.2 Problem definition

Our project will aim at answering the following question:
Can we detect communities of users and can we find musical experts inside each of these communities to provide more accurate and more human recommendations to other users of an online musical streaming service?
The goal will be to find those musical experts and to link them to other users through friend recommendations - based on which community users belongs to.

As the datasets studied contain data for different countries, the model will also be assessed across countries and it will be checked if it is possible to link people from different countries together. The experts and user tastes between countries will be compared to see if they vary in a similar pattern.

1.3 Related work

There are some research which are related to Music Streaming Services (MSS) networks. But mostly the outcomes of those research focus on the recommendation systems of those streaming platforms problems. Our project is very similar to the domain of Recommendation systems, as it aims at finding communities in the social network and giving important friend recommendations to their members.

Application in Business:

Recommendation Systems mainly focus on two areas: link recommendation & object recommendation. Facebook, the most famous social network, focuses on link recommendation where friend recommendations are presented to users. On the other hand, many E-commerce companies such as Amazon emphasize object recommendation where products are recommended to users based on their purchase and browsing history.

Application in Education:

A number of previous studies were devoted to build recommendation systems for education purposes. As an example, a collaborative filtering method was used in a research to recommend documents that will encourage the users to expand their knowledge of a given topic [7]. Similarly, there are also some Kaggle competitions regarding citation networks.

Application in Music Streaming Service:

Several researches regarding music recommendation systems have been conducted which used different recommender algorithms, such as collaborative filtering,content based methods [11][12]. Content based methods are based on similarity between item attributes and collaborative methods calculate similarity between interactions. Collaborative methods work with the interaction matrix that can also be called rating matrix in the rare case when users provide explicit rating of items.

In the domain of Network Science Analytics, there are several popular topics, such as: Centrality, Communities, Link Prediction, Graph Similarity, Graph Sampling, etc. Our project mainly lies in Centrality, Communities Detection.

2 METHODOLOGY

As mentioned, the objective of this project is first to detect communities of users in the network. After the data exploration process and understanding the dataset, different methods were examined in order to find the best separation of the network into communities. The Louvain algorithm was chosen for that purpose and its process is described below. Graph sparsification was also used for visualization and community detection. Finally, musical experts were defined by choosing the nodes that had the highest degree centrality.

2.1. Dataset

The data was originally collected from the music streaming service Deezer in November 2017 by Standford University and provided as open source on https://snap.stanford.edu/data/gemsec-Deezer.html. The data is divided into three datasets all representing mutual friendship networks of users from 3 European countries: Croatia (HR), Hungary (HR) and Romania (RO). The three datasets are three undirected graphs. Nodes represent the users and edges are the mutual friendships.

The number of nodes and edges for each country's network are shown in the table below:

Country	Number of nodes	Number of edges
Romania	41,773	125,826
Croatia	54,573	498,202
Hungary	47,538	222,887

Additional to the friendship network, the data include a file containing the genre preferences of users (where each key is a user/node idea). In each dataset, users could listen to up to 84 distinct genres.

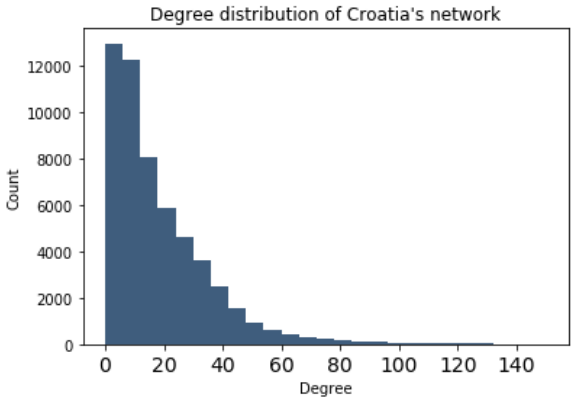
2.2. Data exploration

The first step of the data exploration process was to find the basic statistics of the three datasets. Finding the type of connectivity in the graph, the degree distribution, the average clustering coefficient, etc. gives important information on the structure of the network and on what type of algorithm could best fit the graph.

Country	Romania	Croatia	Hungary
Minimum degree	1	1	1
Maximum degree	112	420	112
Median degree	5	13	8
Mean degree	6.02	18.26	9.38
Number of CCs	1	1	1
Is the graph connected?	Yes	Yes	Yes
Number of triangles in GCC	95373	1992996	294261
Average clustering coefficient	0.09	0.14	0.12

In these statistics it is interesting to note that the giant connected components include 100% of the network's nodes, meaning that the graphs are fully connected. This implies that for every pair of nodes i and j, a path exists between the two nodes.

Moreover, regarding the degrees metrics, the low median and the low mean can both be explained by the degree distribution, which is skewed to the right for every country. For example, the degree distribution of the Croatian network is displayed below:



Also it can be observed that the datasets of different countries are not linked to each other, as they have different indexes of nodes, so analyzing connections across different countries did not seem possible. Considering that, and the fact that all of the dataset had the same statistical properties, it was decided to focus of the Croatia network.

2.3. Louvain algorithm

The Louvain method is an algorithm for detecting communities in networks. It maximizes a modularity score for each community, where the modularity quantifies the quality of an assignment of nodes to communities by evaluating how much more densely connected the nodes within a community are, compared to how connected they would be in a random network.

For this network, it resulted from the Louvain algorithm that highest values of modularity were obtain when the network was divided into 25 communities. In order to get a better understanding of the partition of the network, various metrics were computed on every community. These metrics are shown below:

	internal density	number of edges	average degree	fraction of maximum degree
count	25.000000	25.000000	25.000000	25.000000
mean	0.089196	16063.320000	13.869589	0.376583
std	0.274215	13121.116602	4.175614	0.133089
min	0.001109	3.000000	2.000000	0.000000
25%	0.005767	6446.000000	12.726727	0.359753
50%	0.008270	13495.000000	14.936359	0.409029
75%	0.012269	18849.000000	16.660900	0.454410
max	1.000000	58076.000000	18.268245	0.531095

Community metrics (1)

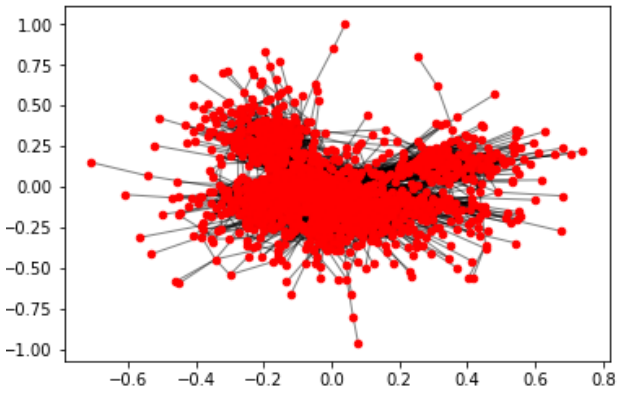
	trangle participation ratio	expansion	conductance	maximum out-degree fraction
count	25.000000	25.000000	25.000000	25.000000
mean	0.836938	1.721515	0.109404	0.819796
std	0.064518	0.694964	0.038091	0.163850
min	0.646911	0.000000	0.000000	0.250000
25%	0.813120	1.449516	0.090302	0.850000
50%	0.830450	1.668639	0.110151	0.857143
75%	0.847911	1.959670	0.131358	0.888889
max	1.000000	3.331839	0.173448	0.933333

Community metrics (2)

The mean average degree and expansion yield that on average nodes in communities have about 13.8 incident edges and only 1.7 of them point to nodes outside the community. The mean conductance of communities is 0.1 which also indicates that the partition is done well – the clusters (communities) are more connected within themselves than with each other. Althow the maximum out degree fraction is quite high which indicates that some rare nodes are still have more connections that point outside the community than inside. The internal density metric shows that within themselves the communities are not very dense – on average only 8% of all possible edges is present.

2.1. Graph visualization

The visualization of the whole network was not feasible due to excessive number of nodes and edges. Instead, focus was brought to visualize the graph of one community of medium size (1 310 nodes and 10 472 edges) of the network. The graph is displayed below:



This image shows that the vast majority of the nodes are concentrated in the center of the graph.

2.2. Graph sparsification

The initial motivation for graph sparsification was to make the visualization less computationally expensive and to compare the results of the Louvain method applied to the sparse graph to the results obtained with the Louvain method applied on the original graph. After choosing local sparsification and the L-spar algorithm was used with sparsification exponent e=0.1. This resulted in a reduction of the number of edges to 54573, which is approximately 10 times smaller than the original graph. Even this important reduction in the number of edges, the computation necessary for the visualization of the graph was still too demanding and time-consuming. Moreover, the results obtained with the Louvain algorithm were significantly different for the graph after sparsification (close to 300 communities instead of 25 for the original graph). The sparsification exponent was tuned from 0.1 to 0.8, but all results were not satisfying – even with the smallest sparsification exponent it was not possible to visualize the graph, and even with the biggest one (e=0.8, results in keeping 322082 out of 498202 edges) the cluster structure was far different from the original one - the network had a big number of small isolated communities instead of several big ones which are connected to each other. So in the end the sparsification was not considered useful for the task.

2.3. Influential nodes and genres distribution

After detecting community of users, the next step is to detect the most important node in every community. These 25 nodes will represent 25 musical experts in the network. Although in practice more experts would be needed, as the network contains more than 40 000 nodes for each country, expert detection was applied on the graph for only 25 nodes. Detection of these musical experts was made by focusing on the degree centrality metric. The idea behind is that a central node is one with many connections. Given that the various communities in the network are quite large, degree centrality seemed a better choice than closeness centrality, which is sensitive to small fluctuations in the structure of the graph and thus not robust enough for a dynamic recommendation system or betweenness centrality, which is computationally demanding and irrelevant for our goal.

As expected, obtained communities correspond to different music preferences. Even though the most popular genres in every community are the most common ones (pop, roc, R&B), communities differ by preference of the less common genres like folk, reggae etc. Moreover, the musical tastes of the selected experts and the other people in the community were explored. The results revealed that the genres preferred by the selected experts correspond to the most preferred genres in the particular community, if the most common genres are excluded. In other words, if the expert listens to folk for example, than this genre will be among the top-5 – top-10 most popular genres in the community. This shows that the results are consistent and nodes with high degree centrality are good representatives of their community.

3 EVALUATION

Our ultimate goal is to find musical experts to link them to other users as friends, so that they can provide music recommendations.

To see if the users of Deezer are satisfied with the expert recommendations, it is easy to conduct a quick survey and submit a questionnaire right after users follow the recommendations by placing it on the Deezer's website or app. It is also possible to use emailing to the users to ask their opinions.

The survey and the questionnaire for evaluation purpose will include some simple rating systems and in the format of Likert Scale. The users of Deezer can express their opinion and the level of satisfaction by simply choosing the number of stars. A five points scale will be used to measure user' agreement, ranging from one extreme to another, for instance a scale of answer choices starting form *Totally disagree* – to – *Totally agree*.

Such methods, are cost-efficient and practical and makes the data collection easy. Online and email surveys allow the users to maintain their anonymity. Mail-in questionnaires also allow for complete anonymity, which maximizes comfort for those answering. This concealment puts respondents at ease and encourages them to answer truthfully.

Digital questionnaires give the best sense of anonymity and privacy. This type of questionnaire is great for all sorts of businesses and subject matter and results in the most honest answers. This will make sure our results will be much more accurate when using this method.

Conclusion

To conclude:

- The proposed method performs well according to statistics and visual observations: the clustering metrics show that detected communities have good quality, plus our observations on musical preferences of detected experts and communities indicate that the results are consistent with the business sense of our task.
- Graph sparsification was not useful for this particular task, as it failed both to simplify the visualization of the full graph and to preserve the clustering structure of the original graph.

A further improvement of this work would be to implement the evaluation process as described above, in order to assess the quality of the recommendations and discover possible ways of advancing the proposed algorithm.

REFERENCES

[1] B. Rozemberczki, R. Davies, R. Sarkar and C.Sutton. GEMSEC: Graph Embedding with Self Clustering. 2018.

[2] Guo, C., & Liu, X. (2015). Automatic Feature Generation on Heterogeneous Graph for Music Recommendation. SIGIR.

[3] Chen, C., Yang, C., Hsia, C., Chen, Y., & Tsai, M. (2016). Music Playlist Recommendation via Preference Embedding. RecSys Posters.

[4] Sangkeun Lee , Sang-il Song , Minsuk Kahng , Dongjoo Lee , Sang-goo Lee, Random walk based entity ranking on graph for multidimensional recommendation, Proceedings of the fifth ACM conference on Recommender systems, October 23-27, 2011, Chicago, Illinois, USA.

[6] S.F.T. Kuan, B.FY. Wu, and W.FJ. Lee, "Finding friend groups in blogosphere," in Advanced Information Networking and Applications; Workshops, 2008. AINAW 2008. 22nd International Conference on, mar. 2008, pp. 1046 – 1050.

[7] Mangina, E. and J. Kilbride, Evaluation of key phrase extraction algorithm and tiling process for a document/resource recommender within e-learning environments. Computers & Education, 2008.

[8] Yiwen Ding,Chang Liu. Exploring drawbacks in music recommendation systems. Bachelor Thesis in Informatics, 2015.

[9] Yading Song, Simon Dixon, Marcus Pearce A survey of music recommendation systems and future perspectives. Conference paper, 2012.

[10] Yajie Hu. A model-based music recommendation system for individual users and implicit user groups. Dissertation, 2014.

[11] Diego Sánchez-Moreno Ana B.Gil González M. DoloresMuñoz VicenteVivian F.López BatistaMaría N.Moreno García. A collaborative filtering method for music recommendation using playing coefficients for artists and users. Expert Systems with Applications Volume 66, 30 December 2016, Pages 234-244

[12] Aaron van den Oord, Sander Dieleman, Benjamin Schrauwen. Deep content-based music recommendation