A dive into the music industry

Sacha Leblanc, Etienne Caquot, Grégoire Mayrhofer, Alexis Mermet

Ecole Polytechnique Fédérale de Lausanne

January 2020

Abstract

We know that there exist relations between artists in the music industry. But how are those artists connected? This paper aims at looking into the various actors of this industry and analyze links between them. For example, we can look at songwriters and see if they write for different artists. We can also look at a specific genre and see if all big artists have worked together or not. If the success of a song depends on the people who worked on it. Or if there exist connections between genres. These are the kind of questions we will answer in the following sections using graph methods and focusing on clustering. We will see that the music industry is a small well-connected world where big names work one with another even though multiple collaborations are rare. We will also discover that studying the collaborations allows sub-genre division for actors that are classified in the same genre on streaming platforms.

1 Introduction

During this project, we wanted to study the relationships between the multiple actors of the music industry. This means we are not only interested in the well known interprets but also the producers and the songwriters. Many artists are superstars but few listeners know the names of all the other persons that worked with an interpret to create a hit song. With our approach, we hope to highlight those unknown actors while understanding the connectivity between them. To tackle this problem, we will use graph theory.

Section 2 explains how we extracted important features based on a database of around 70 000 songs to recover the collaborations that lead to each one of them. In Section 3, we will take a deeper look into the data to understand it before we start the graph analysis. In Section 4, we will try to find out what the industry looks like using multiple graph representations of the collaborations between the artists. Using clustering algorithm we will see how we can divide the actors in their respective genres and also how to regroup each artist of a genre according to his sub-genre. Section 5 highlight the limitations of our analysis and present possible future work, and Section 6 concludes.

2 Data extraction

For our project, we decided to create our data-set. We needed information about artist, writer, producer, label, genre, and popularity associated with a song. First, we found a data-set, created by Zaheen Hamidani in 2018 which can be found on Kaggle[1], giving us a list of songs with the corresponding artist (performer), genre and popularity score.

But this data alone was not enough to perform our project correctly. To gain more information about each song, we used the Genius API[2]. This API enabled us to make HTML requests and gain additional information about each song in the Spotify data-set. In particular, we obtained for each song its songwriters, its producers, and its associated recording labels.

All the work regarding the data extraction can be found in the Jupypter Notebook called *NTDS_project*. We now have the following data structure:

- Song name
- Artist
- Genre
- Popularity
- More Info (Result of the Genius request)

3 Data exploration

With the data being ready, we can start exploring it. First, we started by parsing all the information we could get from the Genius API[2] request made for each song. Unfortunately, for

most of the less known songs we could not get producers or songwriters information with Genius API[2] (Genius is indeed working in the same way as Wikipedia and information can be missing). To face this problem, we decided to put the artist alone in those roles. Here a few numbers about our data:

Songs We have 70681 of them.

Genres A song can have one of the 12 following genre ('Alternative', 'Blues', 'Country', 'Dance', 'Electronic', 'Folk', 'Hip-Hop', 'Indie', 'Pop', 'RB', 'Rap', 'Reggae', 'Reggaeton').

Actors We have 22894 different actors who participated in the creation of the song, those actors are either artist (performer), songwriters or producers.

Popularity Each song is associated with a popularity score, this popularity was in the original Spotify data-set and is an integer between 0 and 100. This popularity score is associated with the number of listening the song has over some time. Here the period corresponds to February 2019, when the data-set was created.

Figure 1 below (bigger in Appendix A), shows the most prolific song writers.

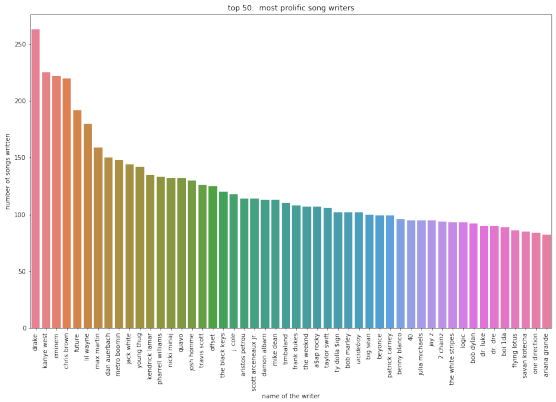


Figure 1: 50 most prolific song writers

This plot shows that a lot of artists are also songwriters. However, we should be careful with our conclusions because we have a lot of missing data. For example, Drake is the biggest songwriter in our data-set but we know that he as been in scandals regarding ghostwriters.

We can also find new figures in this graph. For example, Max Martin appears in the top 10 and after some research, we can find that he is one of the most recognized songwriters in the music industry. He worked for lots of popular artists like Arianna Grande and the Backstreet Boys.

This information tells us that our data grasp information about people behind the songs in addition to the interprets. The study we thus want to lead is possible with the data-set we obtained.

You can find similar analysis in the second Jupyter Notebook called *NTDS_processing_of_the_data*, in which we also study producers, and labels.

The label that is associated with the biggest number of songs is Atlantic Records, and it only represents around 900 songs. This number is really small when we remember that we have a total of 70 000 songs. From this result, we can say that labels are not as important as we thought (note also that lots of big artists have their label). The music industry has changed in the last decade and the labels will not be as useful as we thought to find relationships between the actors in our data-set.

As we are never working with perfect data, this conclusion should not be taken for granted and errors might come from the missing information from the Genius API[2] or our parsing.

Figure 2 (bigger in Appendix B) shows the popularity of the song associated to Atlantic Records.

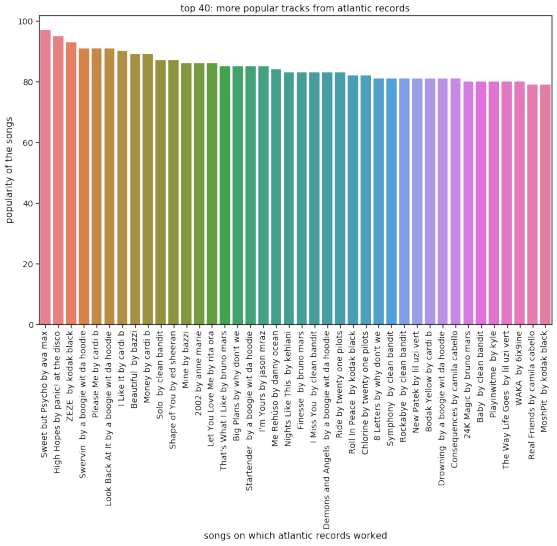


Figure 2: Artist working for Atlantic Records

We can see that the songs published by the label have really good popularity scores. So having a label might not be useful to find connections but it certainly is for an artist to promote their song and gain popularity on streaming platforms like Spotify.

In Figure 3 (bigger in Appendix C) we look at the producer that produced the most songs in our data-set: Rick Rubin. Looking more in-depth at a producer, we can now state some hypotheses about our graph. Producers seem

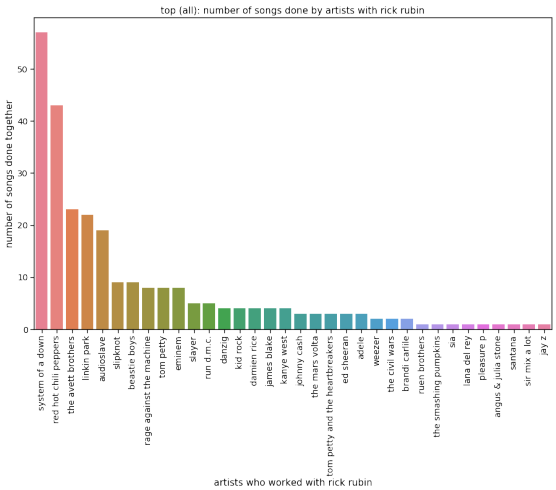


Figure 3: Number of song produced by Rick Rubin for artists

to have a main genre but they also work with artists that perform other styles, these numerous connections between genre might lead to a well-connected graph or at least a giant component with producers being big hubs.

4 Graph analysis

We are now going to dive into our graph approach to study the music industry. We first present to you how we created our adjacency matrix and then talk about what we found out on the data using community detection algorithms.

4.1 Adjacency Matrix

Our adjacency matrix was created so that it represents the collaborations between any actors of the industry. If two persons worked together on a song, we will increment by one the corresponding entry in the matrix (starting with all zeros). An element in the matrix will be an integer representing the number of collaborations between any two actors (see notebook *NTDS_processing_of_the_data*). From this method, we obtained a highly connected matrix that you can see in Figure 4. This matrix contained few zero values which indicate that our network is indeed remarkably connected. We could say that the music industry is such that any actor is reachable from another because collaborations are the norm. One flaw from our approach is that one collaboration is enough to say that two people are linked. So our graph doesn't necessarily show real relationships as friendships or band membership. To solve this problem we defined a second method to create the adjacency matrix. If the number of collaborations between any two persons is superior to 3, then they are connected by a link. Otherwise, they are not connected. With this approach, we obtained a sparse matrix (Figure 5) showing that long term collaborations are still pretty rare in the music industry.

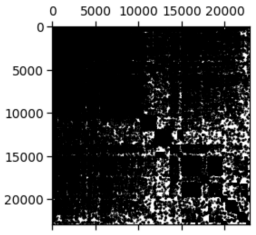


Figure 4: collaborations matrix, black are non-zero values

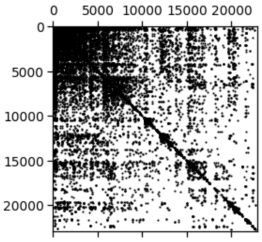


Figure 5: new collaborations matrix, black are non-zero values

In Appendix D, you can see that our first approach leads to a big connected components in the middle of the graph and lots of smaller one all around it. On the other hand, the second method leads to a small number of connected components and lots of actors being all alone in the outside circle of our graph (Appendix E).
Note: the plotting was done only on 20% of the data to avoid lags in the notebook

4.2 Data analysis using clustering algorithms

To find insight in our graphical representation of the data-set, we are going to use the following clustering algorithms.

Spectral clustering uses information from the eigenvalues (spectrum) of special matrices to obtain the community of each node

DBSCAN given a set of points in some space, it groups together points that are closely packed together (points with many nearby neighbors).
The outliers are the nodes found in low-density spaces and are denoted in python with the label -1.

Louvain extract communities from large networks by optimization of the modularity value which "measures the strength of division of a network into modules".

First of all, we used these three methods on two distinct adjacency matrix corresponding to the ones you can see in Figure 4 and Figure 5. Spectral clustering on the first matrix yields that we indeed have one big cluster where any two persons are connected (Figure 6). This cluster corresponds to the biggest artists in the industry. Then, the algorithm does not work well on the outside ring because many elements are not connected at all. From this result, we obtained the insight that the industry is composed of two small worlds. One very connected corresponding to what seems to be the popular music and one very noisy corresponding to the more independent music scene and the small artists.

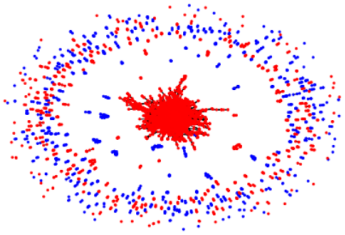


Figure 6: spectral clustering applied to the first adjacency matrix

Applying spectral clustering to the second matrix, we can see that the algorithm is not of any help. It finds some good cluster in the middle of the graph but is unable to find one for all the elements in the outside ring that are not connected at all (Figure 7).

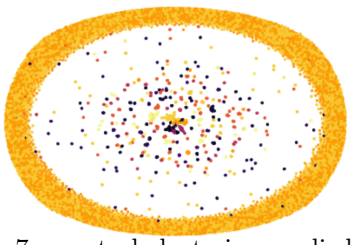


Figure 7: spectral clustering applied to the second adjacency matrix

We then applied DBSCAN and, as you can see in the notebook *NTDS_processing_of_the_data*, it didn't gave us interesting results in both cases. Finally, we applied the Louvain method for community detection. This time, the found communities lead to interesting results as you can see in Appendix F and Appendix G. For the first, matrix we found, for example, a cluster (in cyan) corresponding to the majority of the popular Rap/Hip-Hop music in today's standards. It contains artists like Drake, Future and Travis Scott. We also have in light green and dark purple "pop" clusters with artists like Arianna Grande or Chris Brown. Finally, we can observe in blue a cluster corresponding to the more Old School Rap with the presence of Eminem or Dr.Dre. From this partitioning, it seems like Louvain achieved to find genre-related clusters. We could then suppose that even though the actors in the musical worlds collaborate a lot, independently of their genre preference, they still are more active with artists doing the same type of music. We can obtain the same insight with the second matrix (Appendix G) where the grey cluster corresponds to popular rap and pop music(Ed Sheeran, Drake, Taylor Swift). The cyan one to RNB music and the dark blue to actual and former members of Shady records (Eminem's label). We can also see in light green a cluster corresponding to artists who worked with Rick Rubin (songwriters and producers are clearly hubs). Note the presence in both graphs the existence of a Latino music clusters (In grey on the right in Appendix F and in purple on the top corner in Appendix G).

From this analysis, it seems like clustering methods allows to group artists based on the genre they worked on the most. This seems logical since it is more likely for artists to work more with people having the same musical vision as them than with any other. We can also observe some hiatus in the rap music genre. From this observation, we decided to study if the Louvain method allows us to do in addition to genre detection, sub-genre detection (for example old school rap, cloud rap, gangsta rap, ...).

To assess this hypothesis, we decided to look only into Rap/Hip-Hop songs by creating a new adjacency from only such songs. In Appendix H, we can see that our intuition is true and that the study of collaborations provides sub-genre division of the different artists working on Rap/Hip-Hop songs (even though we can't give proper names to these clusters). In blue, you have a cluster corresponding to popular rap music from rappers like Drake. In green, you still have old school rap with Shady records members. In yellow you can find the rap similar to Kendrick Lamar. Or in Cyan a cluster with Kanye West and Kid Cudi.

Finally, we saw at the start of this section that the majority of popular artists can be found in the same connected components in the center of the graph (Figure 6). Thus we took interest in looking only at popular artists. We first created an adjacency matrix using only information about popular songs (at least a popularity of 70 at the time of pulling) and obtained the graph in Appendix I. This time, the outside ring nearly disappeared (see in the notebook *NTDS_processing_of_the_data*). The majority of popular artists are in the middle of the graph. From this result, we can deduct that col-

laboration help song to get popular. Also, even if popular artists are there as expected, some less known artists still find a place in this graph (usually) if they collaborated at least once with a big name in the industry. We could then suppose that achieving success in the music industry is hard and songs that casual listeners access to are the product of a small group of people (our matrix here contains only 5171 artists from the 22894 in the full data-set).

5 Discussion

5.1 Limitations

Now that we have presented our research, let's discuss the results in this Section. First, we would like to highlight that the results of this analysis might not reflect the reality of the music industry. Indeed we could only work on 70 000 songs which is a fraction of the total available songs on Spotify (50 Million +), as a result, all our clustering and component analysis in Section 4 is probably different than the real-world picture. If we were able to have more data, the more likely there would be links between the different actors and thereby we would have different results for the clustering methods.

However, we think that even with more data the results would be quite similar. There are lots of little artists who work independently and who would never be part of the giant component we could observe using Spectral clustering. We could also only create the adjacency matrix where we want a certain number of collaborations to link two actors (see Subsection 4.1). This can be reinforce considering that the Louvain method was able to find sub-genres in a given genre, meaning that even with small data-set we could find strong connections between similar actors. Given the poor data, this is quite impressive.

Another important point to notice is the number of songs for which we could not have more information with the Genius API. Since the Genius platform is collaborative, it is more likely that well-known artists have information such as their label, producers and songwriters whereas the smaller artist will probably won't have all of the information. This problem has certainly reinforced the connectivity between the giant component mainly composed of big actors of the industry for which the links between them were easy to make.

5.2 Going further

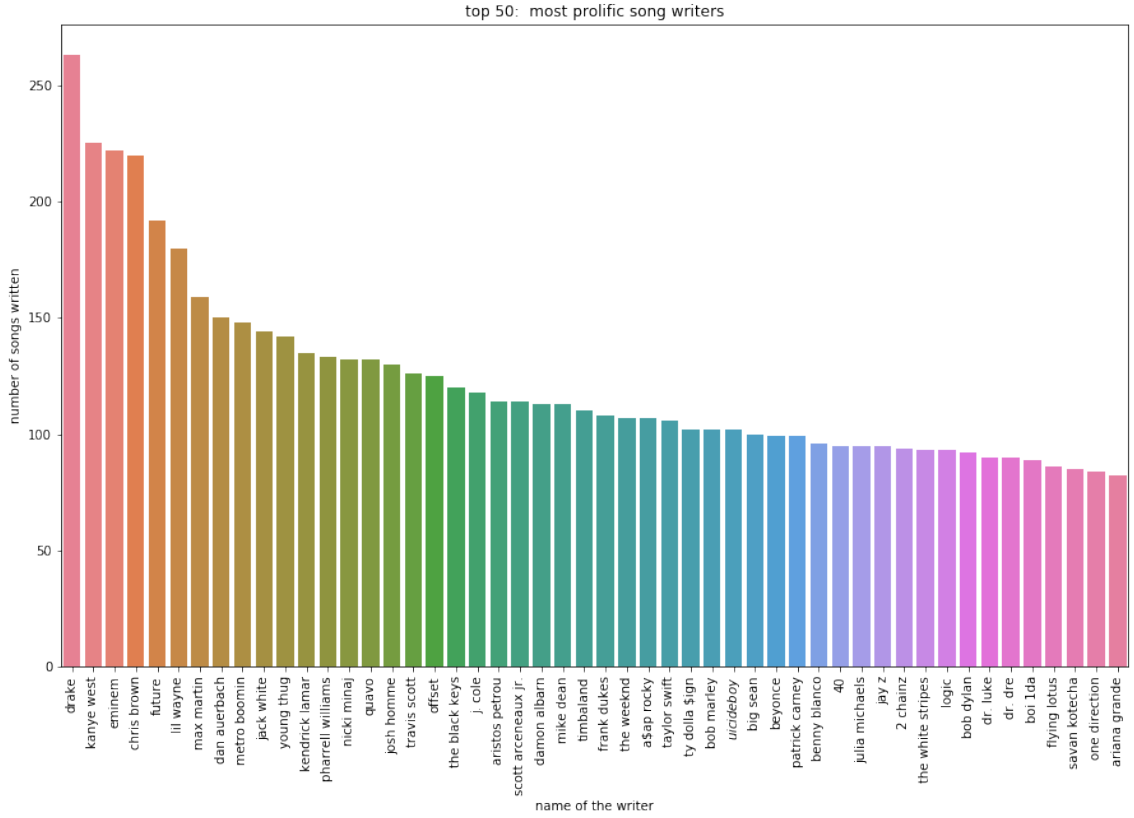
The next step of the work would be to extend the data-set as much as possible. There are various libraries available to communicate with the Spotify Web API. This might be really hard work to do since Spotify seems to limit the number of requests. Our analysis was already computationally intense and having more data would mean that we would need to increase our computational power, which is hard without having access to a server.

Regarding the analysis itself, it would be interesting to extend the analysis to several years and observe the evolution of graphs and connections over time. We could also try to focus on a specific actor or a specific label looking with whom connections are made and if we can find any structure around this actor of the music industry.

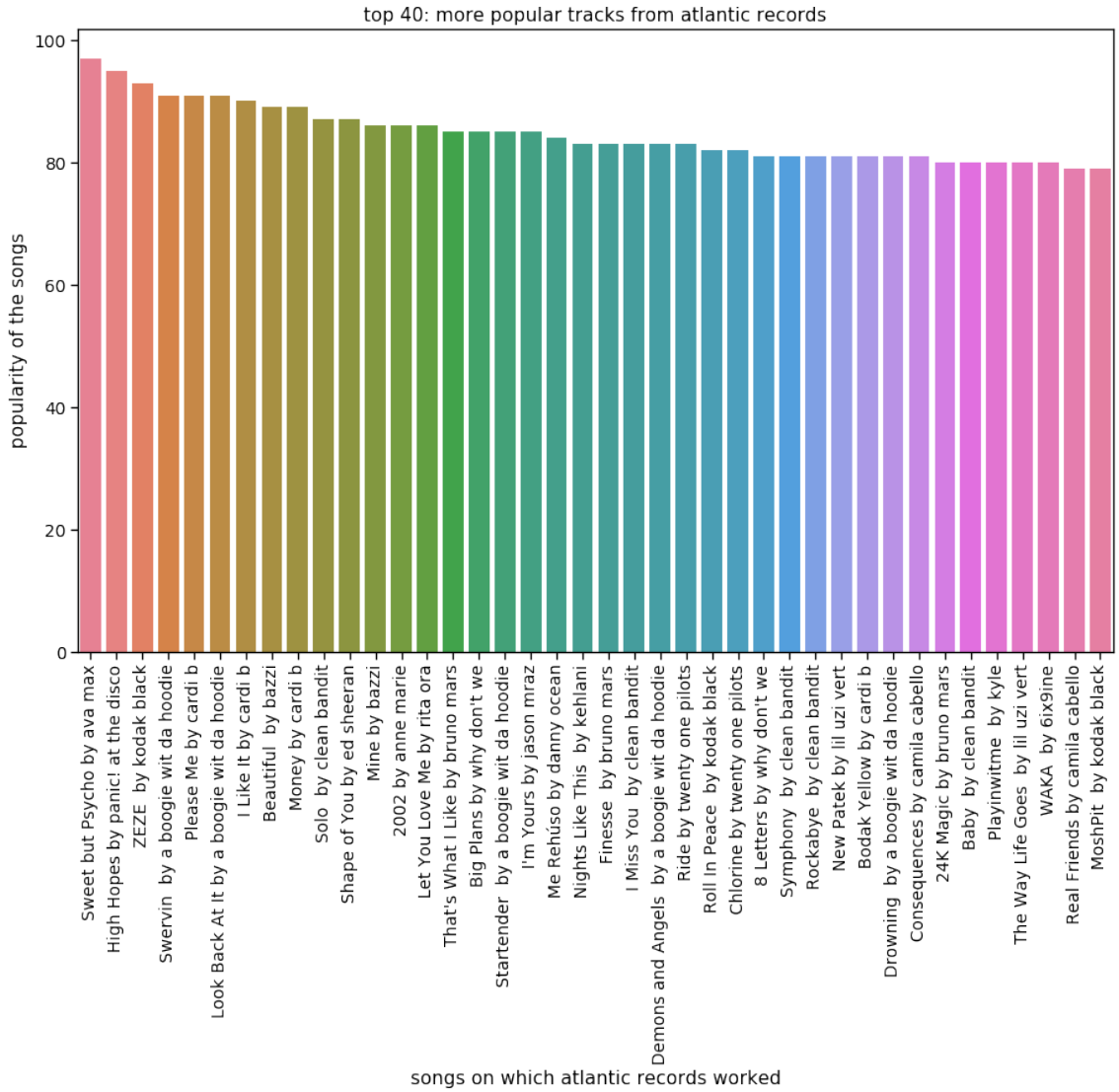
6 Conclusion

This work has confirmed the prejudices we have from the music industry: is it a small and well-connected world where a small group of artists is responsible for the majority of successful songs. All big names seem to be highly connected and it's hard for one artist to stand out compared to all the others. It seems pretty hard to achieve success without knowing people that matter in the music industry since the small independent actors don't link to the big ones. An artist which is new is more likely to get successful by doing a collaboration with an already established artist in the same genre, thus gaining popularity which will play a role in his future songs. Graph clustering was a good approach to study this industry which pretty much works like a small world. Because our graph was not fully connected, we couldn't compute lots of measurements that would have helped us to understand better the connection between the big name of the industry (like the diameter of our graph). Despite these problems, graph clustering allowed us to do some sub-genre recognition which could be useful on a platform like Spotify.

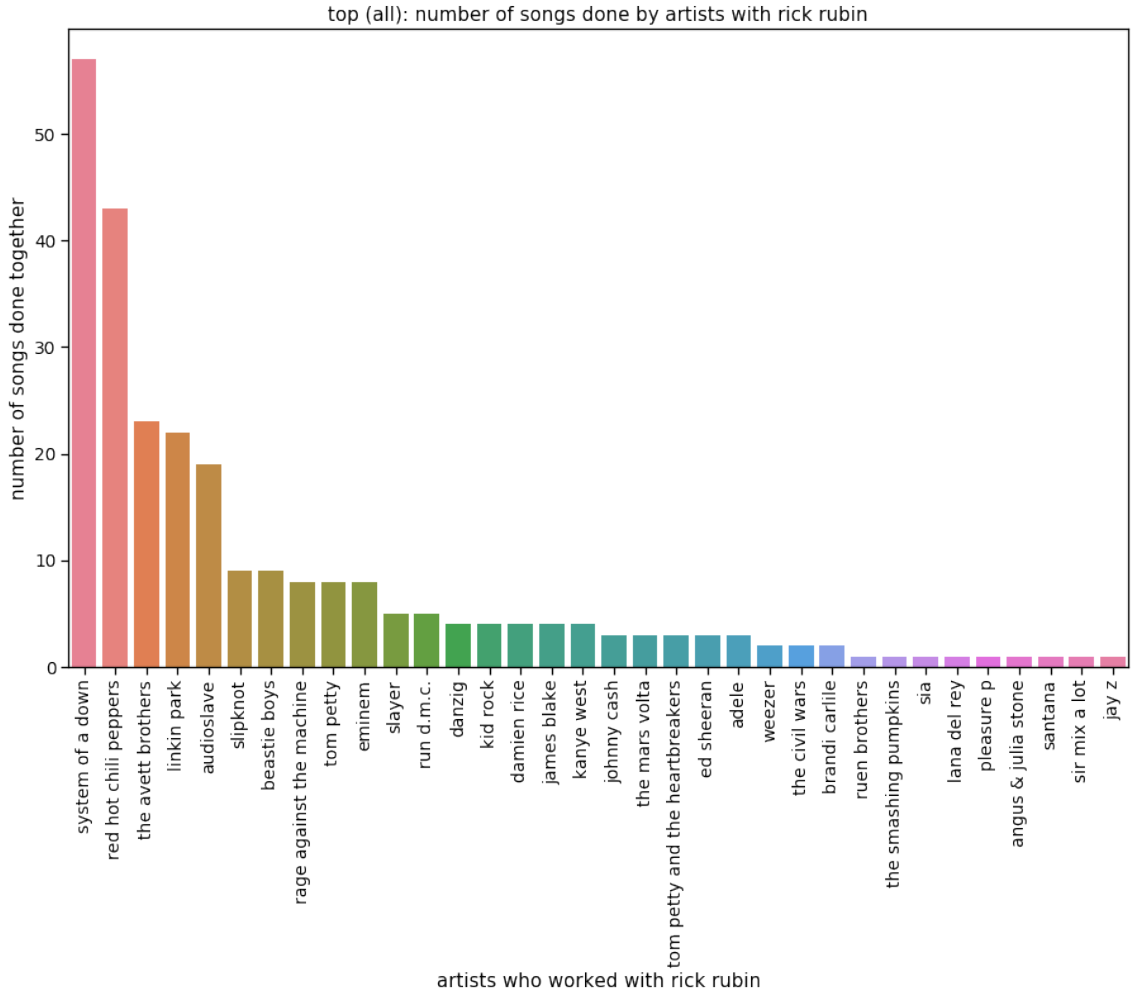
A Most prolific song writers



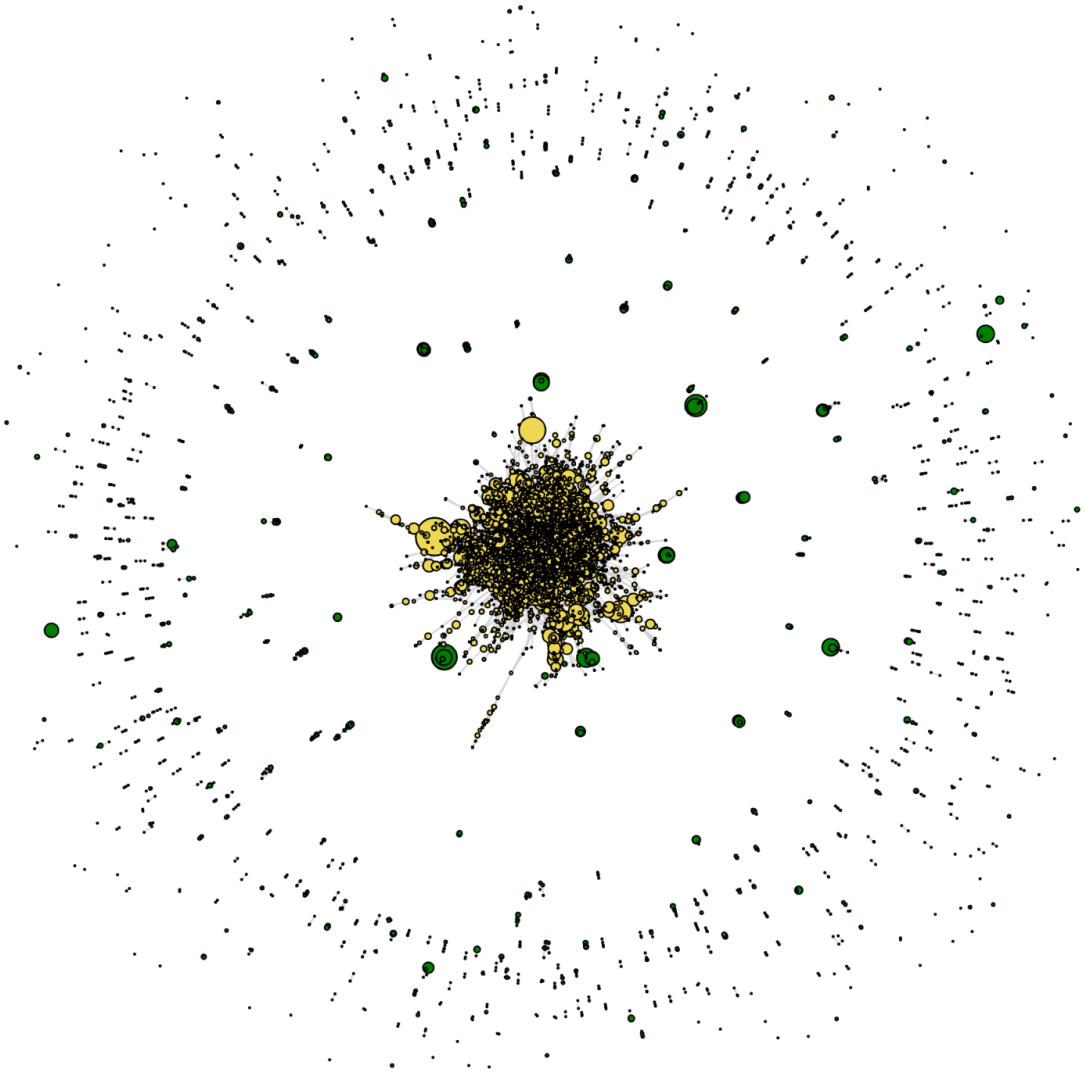
B Most popular songs under the label Atlantic Records



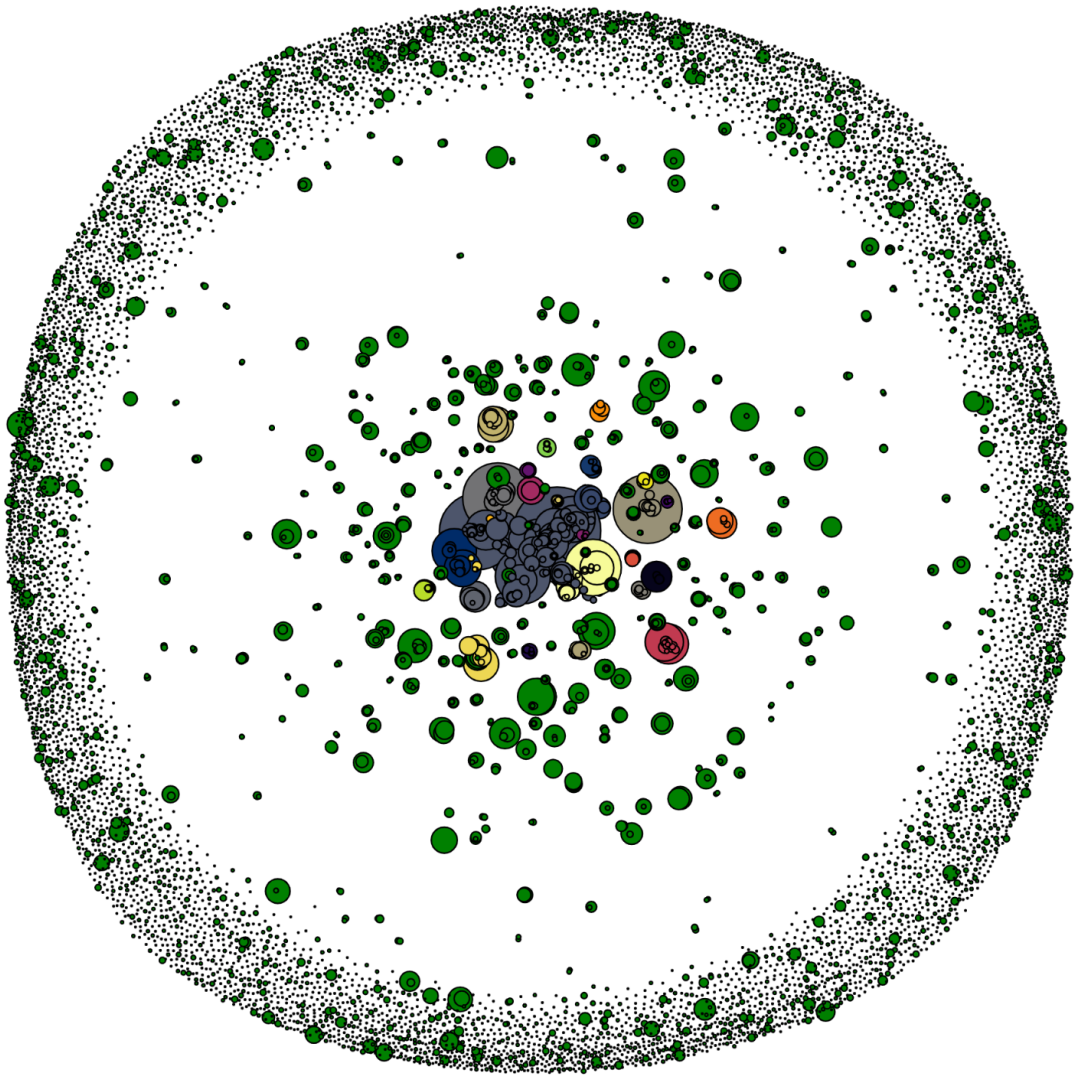
C Number of songs produced by Rick Rubin per artist



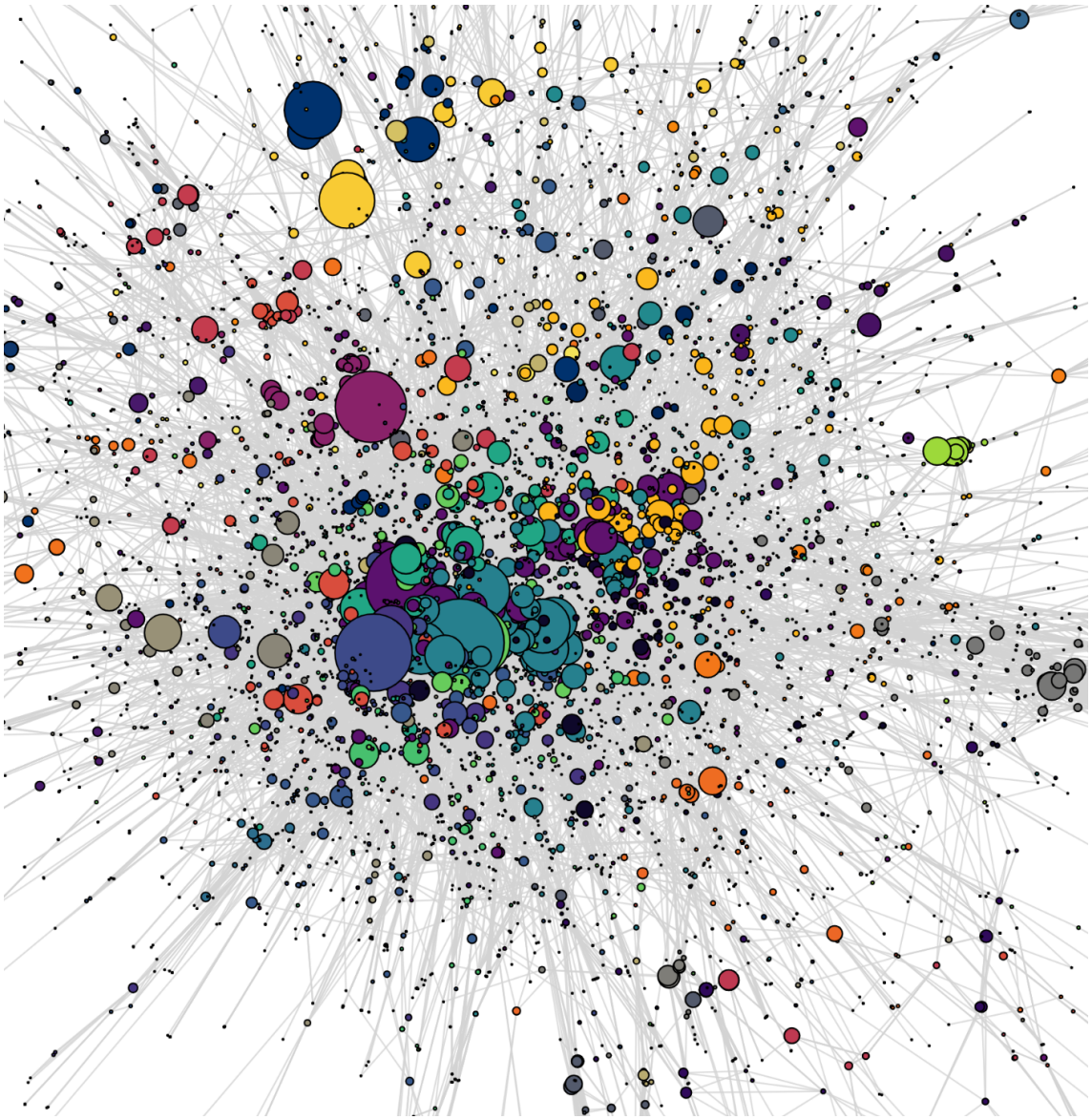
D Graph corresponding to the adjacency matrix containing all the collaborations. Coloring obtained from connected components of size > 20



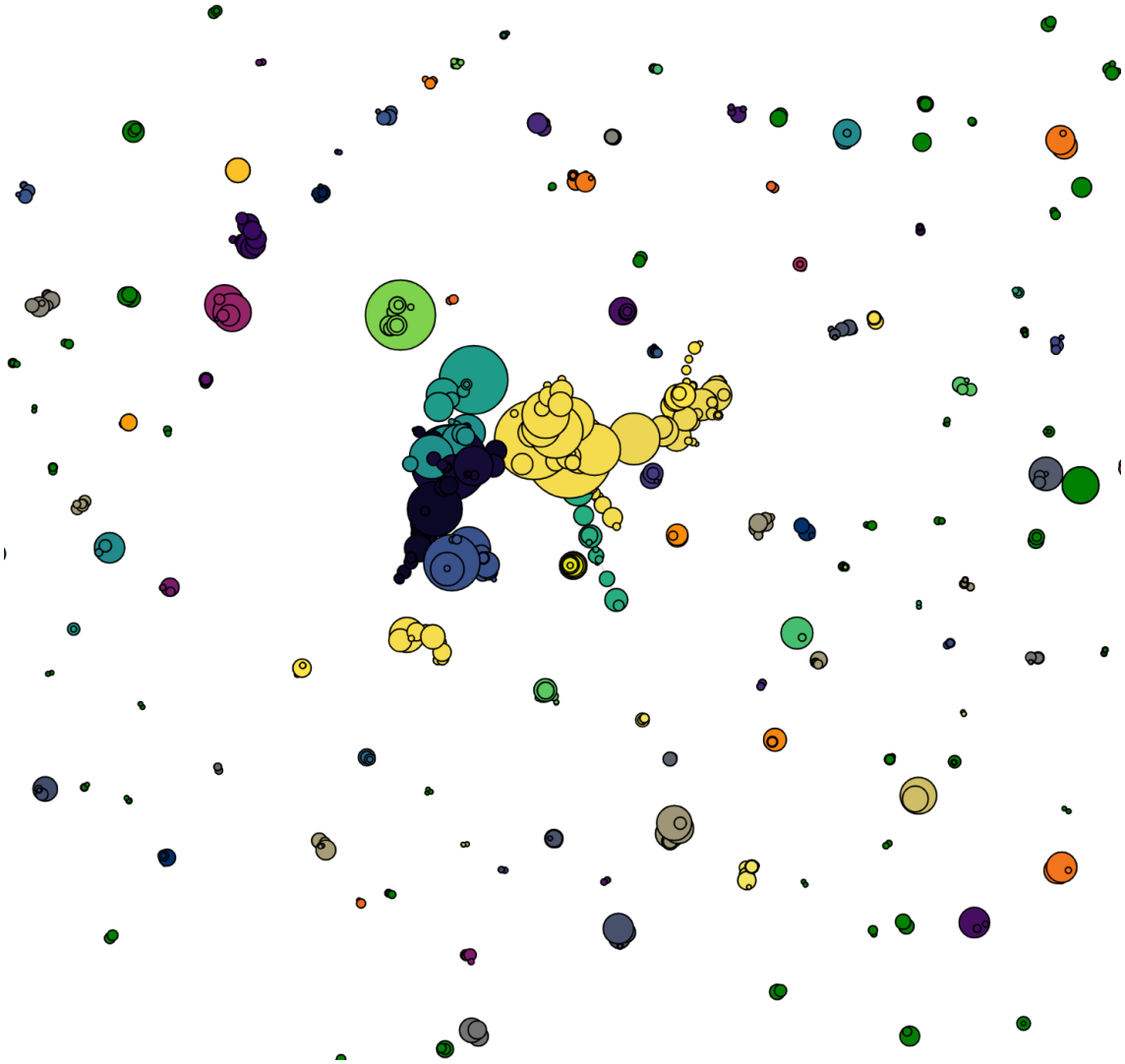
E Graph corresponding to the adjacency matrix containing only links between people that collaborated more than 3 times. Coloring obtained from connected components of size > 7



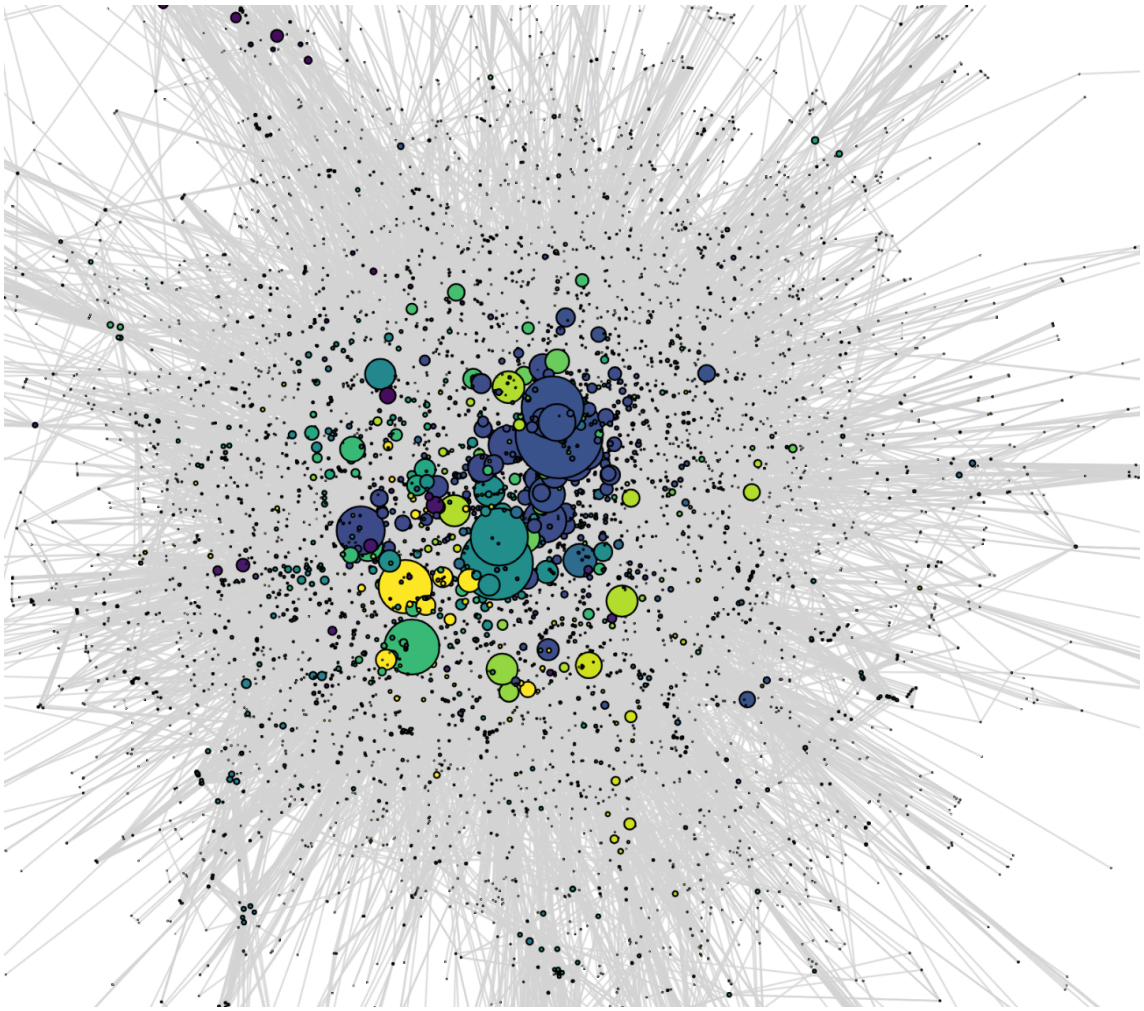
F sneak pick into the Louvain method applied to the adjacency matrix containing all collaborations. Coloring obtained from clustering



G sneak pick into the Louvain method applied to the adjacency matrix containing only multiple collaborations. Coloring obtained for cluster of size ≥ 5



H sneak pick into a graph representing the rap/hip-hop world. Coloring obtained using the Louvain method



I sneak pick into a graph representing the popular artists. Coloring obtained using the Louvain method

References

[1] "Kaggle data-set." https://www.kaggle.com/zaheenhamidani/ultimate-spotify-tracks-db.

[2] "Genius API." https://docs.genius.com.