Class Project for
CSCI 5352: Network Analysis and Modeling
Investigation into the Music Network of SoundCloud

Kansuke Ikehara

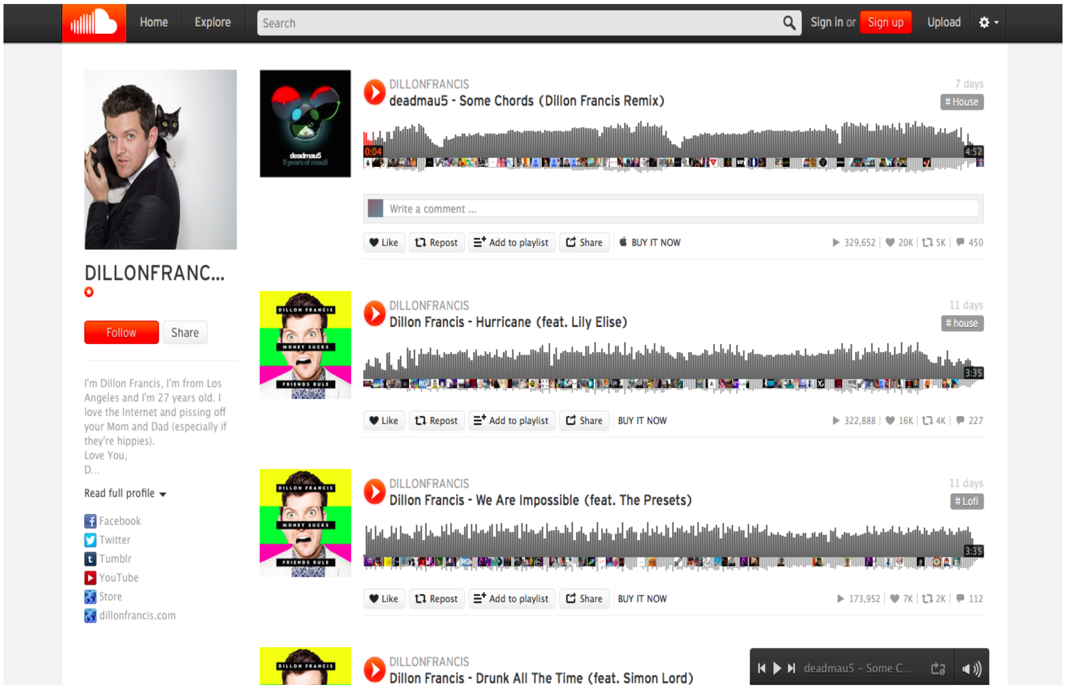
December 14, 2014

Abstract

In this paper we investigate the music network of SoundCloud, the music sharing social network service. First, data of 5000 users' favorite tracks are sampled via SoundCloud API. The sampled data form a bipartite network of users and musics and the network is then one-mode projected onto a music network in which each edge between two vertices (tracks) corresponds to how many users "like" the two tracks, resulting in a weighted network. The resulting network is thresholded so that it becomes compact enough to be applied a community detection algorithm for analyzing the emergence of music genres among the network. We observe that the algorithm finds 9 communities, though several of these communities do not seem distinct from others. We also observe that dominant music genres in the network are: House/Electro and Hip-Hop/Rap. Finally, we conclude this study with discussion of *core-periphery* structure within the observed network.

1 Background

SoundCloud has two distinct faces to two different groups of people. One is as a music sharing platform for artists who are motivated to share their musics with people around the world. Another is a social network service for music listeners where not only users can listen to musics for free via streaming, but also can make a comment on a music, like artists, tracks and playlists of other users and follow their favorite producers or other users. Fig.1 shows a screen shot of a SoundCloud's web page.



I'm Dillon Francis, I'm from Los Angeles and I'm 27 years old. I love the Internet and pissing off your Mom and Dad (especially if they're hippies).
Love You,
D...

DILLONFRANCIS
deadmau5 - Some Chords (Dillon Francis Remix)
7 days
House
329,652 | 20K | 5K | 450

DILLONFRANCIS
Dillon Francis - Hurricane (feat. Lily Elise)
11 days
# house
322,888 | 16K | 4K | 227

DILLONFRANCIS
Dillon Francis - We Are Impossible (feat. The Presets)
11 days
Lofi
173,952 | 7K | 2K | 112

DILLONFRANCIS
Dillon Francis - Drunk All The Time (feat. Simon Lord)

Figure 1: a screen shot of SoundCloud's web page

The question arises when considering SoundCloud as an enormous network data set: **When we do one-mode projection of a user-music bipartite network, in which edges are "likes", onto a music network, does the network exhibit an assortativity based on their musical characteristic or *genre*?** Lambiotte and Auslood [1] studied the bipartite network of users and their music libraries, whose data was downloaded from *audioscrobbler.com* in January 2005. Their study shows that there is an emergence of music groups in the music network which is one-mode projected from the bipartite network. Does the same thing happen in the music network of SoundCloud, which is much newer and bigger than the data analyzed in [1]? And what kinds of interesting observation will we get through the investigation? In this class project, we try to answer those questions.

2 Analyzed Data

To analyze data, we first need to earn the data in some way. Fortunately, SoundCloud provides developers an API via which they can collect various kinds of data either as an XML or JSON file. The data includes: SoundCloud user's information such as a user ID, user name, favorite tracks, playlists etc. and uploaded music's information such as a track ID, track name, tags and so forth. In this study we have randomly sampled 5000 users' favorite musics as JSON format files, which count up to 252485 tracks, via API. Sampled users have at least 20 and at maximum 200 musics which they have "liked". Sampled data then form a bipartite network of users and tracks. The reason for this filtering users based on how many musics they have liked is to give a music network, which is going to be discussed in the following chapter, a fair amount of edge weights among tracks.

3 Analysis methods

3.1 One-mode Projection

One-mode projection of a bipartite network is to project the network onto an either side of two groups. In this study, we project the bipartite network of users and musics onto a music group. If two musics are liked by one user, in the resulting music network, those two tracks have an edge whose weight is 1. If they have more users liking them, the weight becomes **how many users have liked them**. Therefore, the music network is an weighted undirected graph. We define that edge weights in this network represent the degree to which two musics are similar in terms of a music taste or *genre* based on the assumption that the people' music preference is highly dependent upon a music's genre. For example, young music listeners might tend to prefer trendy pop musics or club musics such as EDM or intense musics such as Metal, Dubstep, etc. whereas aged people might like relatively relaxing musics such as classics (Sorry for a very extreme example).

The resulting music network contains 252485 vertices and 22360461 weighted edges. Fig.2 shows the distribution of edge weight of the network. From this figure, we can observe that edges whose weight is 1 account for about 98% of the total edges and the shape of the distribution looks somewhat like a power-law distribution above *weight* = 2.

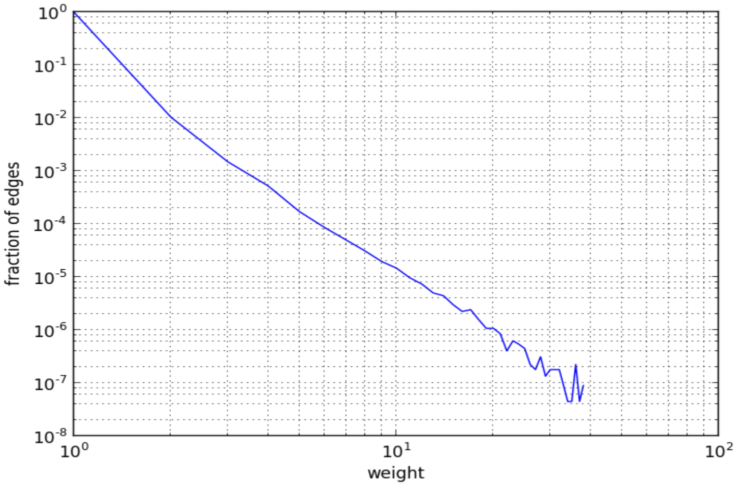


Figure 2: the edge weight distribution

3.2 Thresholding edges

The resulting music network has an enormous amount of edges and the majority of them have weight of 1, which makes the analysis difficult and hides the underlying structure of the network. Therefore, we need to refine the network by thresholding edges based on an edge weight. In this study, we chose threshold value $t = 11$. By thresholding, some parts of the network are separated and result in some small components. Small components having only one vertex are omitted in this process. Fig.3 depicts the network thresholded with $t = 11$. The width and intensity of color of an edge are proportional to the weight of the edge. In Fig.3, two small components and one giant component are observed. The number of vertices is 199 and the number of edges is 735.

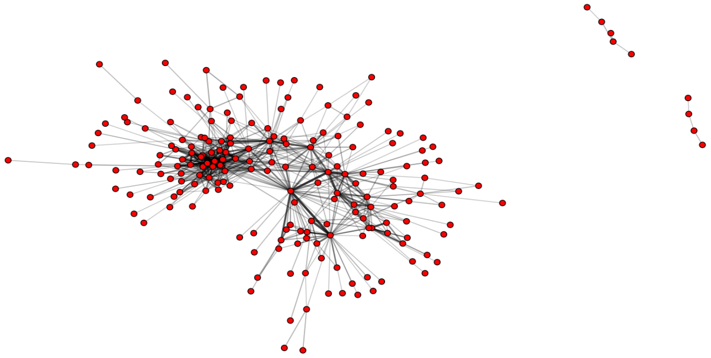


Figure 3: thresholded graph

3.3 Community Detection

If we carefully look at the thresholded network in Fig.3, there seems to be a community structure in the largest component: the one with densely connected vertices in the left side and the another with relatively sparsely connected nodes in the right side. We need, however, a quantitative analytical tool in order to validate the community structure in the resulting network. The community detection algorithm provides us such a tool and lets us find the underlying community of the network. The algorithm we use in this study is the weighted edge version of the agglomerative community detection algorithm proposed by Clauset *et el.* [2] (see [2] for details about the algorithm). The analysis regarding the weighted edge network are well explained by Newman [3].

Let's define some equations for the algorithm. The adjacency matrix of the music network is:

$$A_{ij} = \begin{cases} w \geq 12 & \text{if vertices } i \text{ and } j \text{ are connected,} \\ 0 & \text{otherwise.} \end{cases} \tag{1}$$

The degree of a vertex i in this network is defined as the total value of weights of edges attached to the vertex, namely:

$$k_i = \sum_j A_{ij}. \tag{2}$$

The definition of the modularity Q is:

$$Q = \frac{1}{2m} \sum_{ij} \left[A_{ij} - \frac{k_i k_j}{2m} \right] \delta(c_i, c_j), \tag{3}$$

where m is the total value of weights of edges in the entire network. The community detection algorithm basically tries to find a division of vertices with the optimal (or sub-optimal) modularity Q. The agglomerative algorithm gains ΔQ when merging two communities and that is:

$$\Delta Q = 2(e_{ij} - a_i a_j), \tag{4}$$

where

$$e_{ij} = \frac{1}{2m} \sum_{v,w} A_{vw} \delta(c_v, i) \delta(c_w, j), \tag{5}$$

and

$$a_i = \frac{1}{2m} \sum_v k_v \delta(c_v, j). \tag{6}$$

Remember that m is the total value of weights of edges in the entire network.

4 Results

When we apply the agglomerative algorithm to the thresholded network, it finds 9 partitions of vertices as shown in Fig.4. Fig.5 shows the change of the modularity Q over merging. The maximum modularity Q is 0.464079, which means the algorithm found the obvious community structure within the network. Table 1 shows top 20 of genre or tag attached to tracks. Tracks in group 6 and 8 do not have any genre information or tags.

5 Discussion

From Fig.4, we can observe three components of the music network, one containing the largest number of vertices and the others consisting of few vertices. Although the largest component consists of 7 communities detected by the algorithm we used, the majority of vertices can be incorporated into 3 communities: group 1, group 2 and group 4. Let's first have a look at the group1.

The top 20 genre or tag information attached to the tracks in the group 1 are: hiphop, ovo(which comes from the name of the record label "OVO Sound" founded by rapper "Drake"), drake, r&b, etc. Given the number of tracks in group 1(66 tracks), it might be a weak assertion that the genre of musics in the group 1 as a whole is HipHop/Rap/R&B music based on the number of genre or tag information. However, quite a few people uploading their musics do not attach any

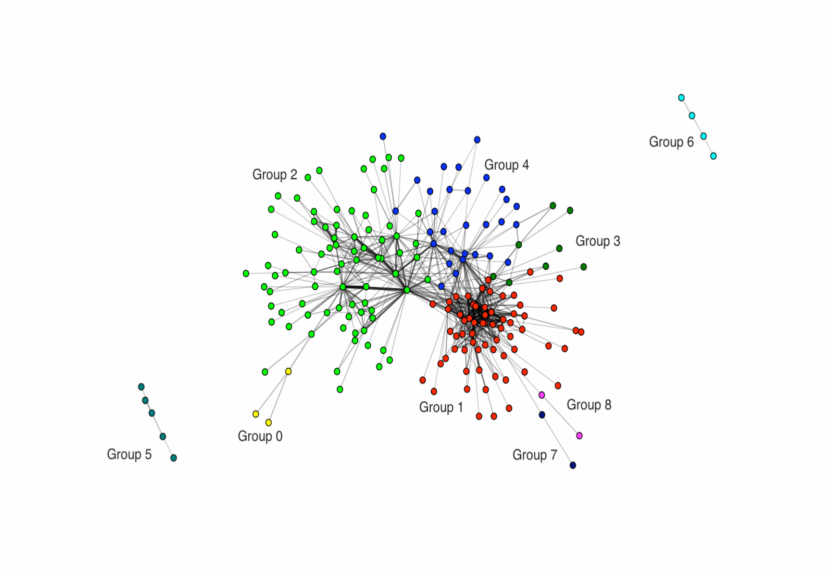


Figure 4: detected community structure

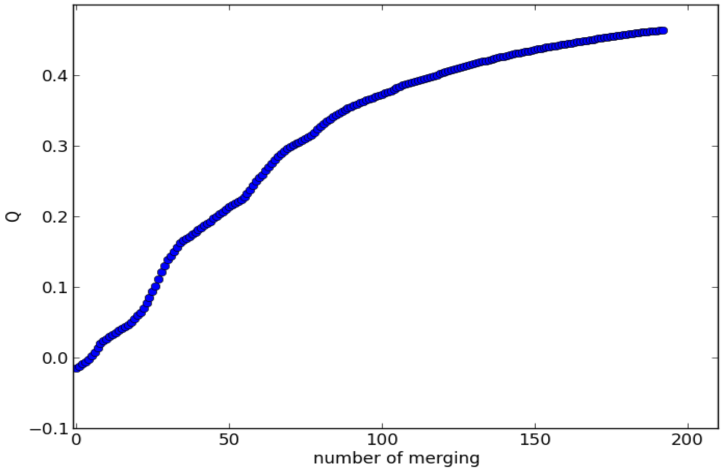
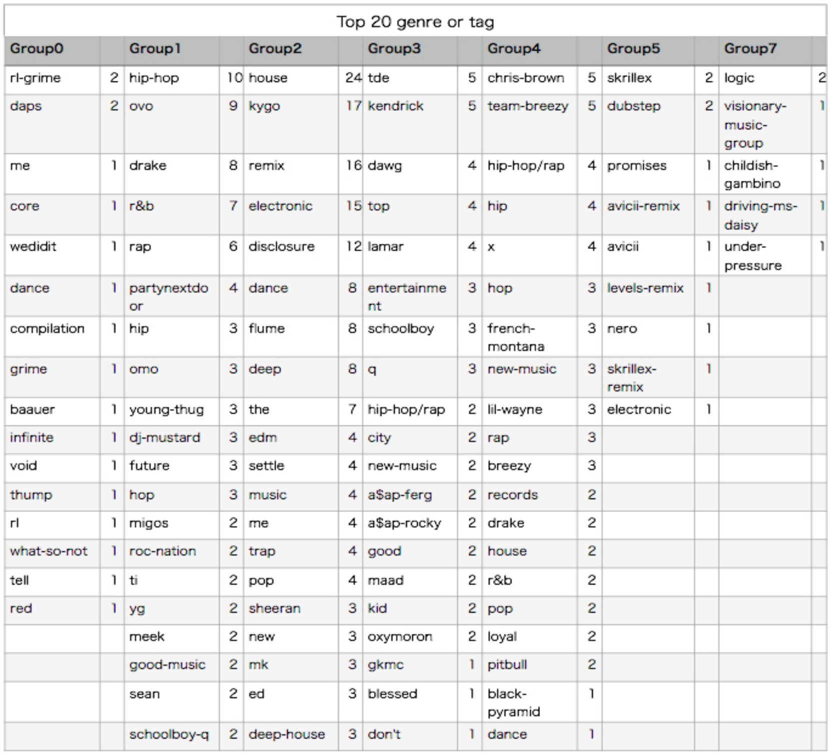


Figure 5: the modularity Q over merging

genre or tag information following the traditional music genre classification such as HipHop, R&B and so forth even though their tracks are considered belonging to those genres. They rather tend to attach genre or tags related to their musical concepts or their names themselves or completely unrelated silly tags to the uploaded musics. Since this tendency is seen in other groups, we, in this study, assume that the music genre of the tracks in a group can be represented by the top 5 of the most frequent genre or tag information. Another thing worth noting regarding the group 1 is the density of connections. Musics in the group 1, especially at the *core* of the group have a far denser connections among them than those in other partitions. In other words, the probability for those musics in the group to be liked together by individuals is higher than those in other groups. This could be explained by the hypothesis as follows: those musics are popular, thus have been played more often than others, leading to a strong likelihood to be liked together.



Top 20 genre or tag													
Group0		**Group1**		**Group2**		**Group3**		**Group4**		**Group5**		**Group7**	
rl-grime	2	hip-hop	10	house	24	tde	5	chris-brown	5	skrillex	2	logic	2
daps	2	ovo	9	kygo	17	kendrick	5	team-breezy	5	dubstep	2	visionary-music-group	1
me	1	drake	8	remix	16	dawg	4	hip-hop/rap	4	promises	1	childish-gambino	1
core	1	r&b	7	electronic	15	top	4	hip	4	avicii-remix	1	driving-ms-daisy	1
wedidit	1	rap	6	disclosure	12	lamar	4	x	4	avicii	1	under-pressure	1
dance	1	partynextdoor	4	dance	8	entertainment	3	hop	3	levels-remix	1		
compilation	1	hip	3	flume	8	schoolboy	3	french-montana	3	nero	1		
grime	1	omo	3	deep	8	q	3	new-music	3	skrillex-remix	1		
baauer	1	young-thug	3	the	7	hip-hop/rap	2	lil-wayne	3	electronic	1		
infinite	1	dj-mustard	3	edm	4	city	2	rap	3				
void	1	future	3	settle	4	new-music	2	breezy	3				
thump	1	hop	3	music	4	a$ap-ferg	2	records	2				
rl	1	migos	2	me	4	a$ap-rocky	2	drake	2				
what-so-not	1	roc-nation	2	trap	4	good	2	house	2				
tell	1	ti	2	pop	4	maad	2	r&b	2				
red	1	yg	2	sheeran	3	kid	2	pop	2				
		meek	2	new	3	oxymoron	2	loyal	2				
		good-music	2	mk	3	gkmc	1	pitbull	2				
		sean	2	ed	3	blessed	1	black-pyramid	1				
		schoolboy-q	2	deep-house	3	don't	1	dance	1				

Table 1: Top 20 genre or tag.

To prove the hypothesis explained above, we first investigate play counts and the number of ”likes” a music has obtained.

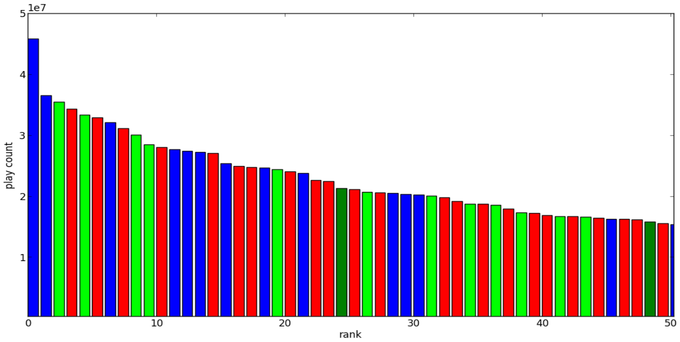


Figure 6: Top 50 of play counts.
Colors correspond to those in group partitions

Figures 6 and 7 show top 50 of play-counts of tracks and that of like-counts, respectively. Colors in figures above correspond to those in group partitions. The track which ranks as the highest in both rankings is *All of me* by John Legend. Worth noting here is that in top 50 of play-counts, tracks in the group 4 (blue) rank at high positions, while in top 50 of like-counts tracks in the group 1 (red) relatively dominate the higher rankings.

Fig.8 shows the results of linear regression of scattered plots of each group. X-axis represents play-counts and Y-axis corresponds to the number of likes each music has obtained. From this figure, the gradient of line of group 1 is larger than others, without groups 7 and 8 being excluded. This finding implies that musics

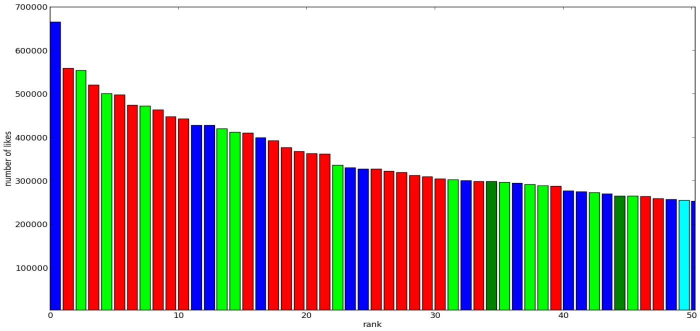


Figure 7: Top 50 of like counts.
Colors correspond to those in group partitions

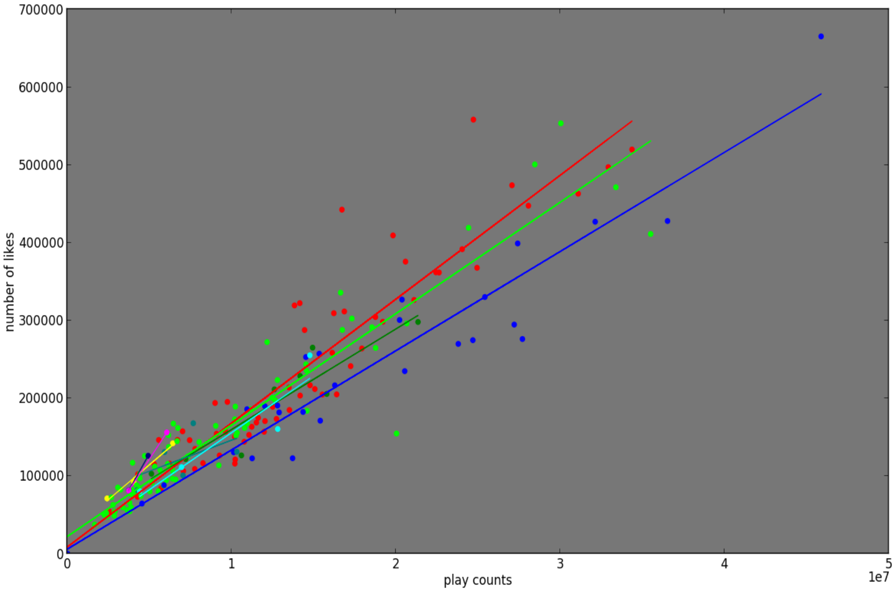


Figure 8: Linear regression of each group

in the group 1 have a stronger likelihood to be liked by the listeners compared to those in other partitions if they all have the same number of play counts. Therefore, the dense connections within group 1 is not because the tracks in it are popular, but rather because they are somehow liked more often than others with the same order of play-counts

Next we discuss the group 2 and 4 from the result. The top 20 genre or tag in the group 2, unlike group 1 containing a number of HipHop/R&B musics, consists of ”house”,”kygo”, a name of a Norwegian DJ, ”remix”, ”electronic” and so on, all of which imply that musics in this group belong to House/Electronic genre. From the network structure perspective, group 1 forms relatively sparse connections within the partition and also exhibits core-periphery structure similar to group 1. Group 4 contains genre or tag information such as ”chris-brown”, an R&B/HipHop singer, ”team-breezy”, a terminology describing the enthusiastic supporters of Chris Brown, ”hip-hop/rap” and so on. Therefore, based on the same idea with group 1, we also assume that group 4 also belongs to HipHop/R&B class.

Lastly, we briefly discuss the *core-periphery* structure within the observed network. In Fig.4, a number of vertices jut up from the central part of communities, namely, the *core*. Especially in group 1 nodes densely connected by edges with strong color and thickness are present at the core of the partition and those connected to the core by thin edges with weak color are located at the periphery of the group.

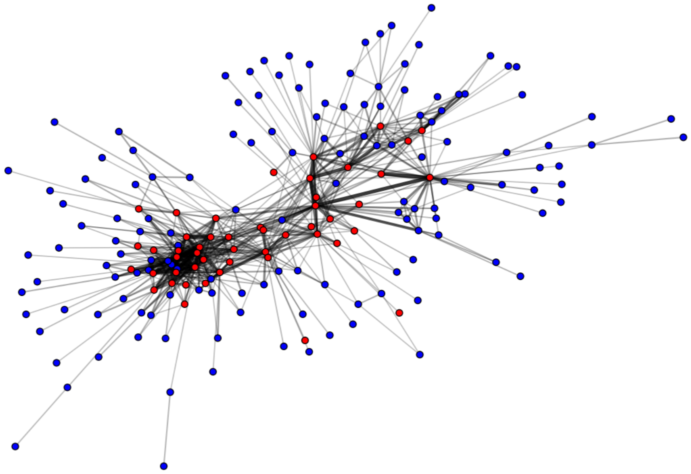


Figure 9: red-colored vertices correspond to tracks in the top 50 of like-count

Fig.9 shows the network in which red-colored vertices are the tracks in the top 50 of like-count ranking. As in this figure, with few exceptions, highly liked tracks are residing in the central of the partition and others are at the outskirt. This finding implies that there may be a correlation between a location of a node within the network and countable meta data attached to the node such as like-count. That is, the more central a track is located at, the more likes (thus more play-counts) it has earned. Another conjecture regarding the *core-periphery* structure is that possibly thresholding edges based on their weight might peel off a *layer* of peripheral vertices, resulting in the spiky outskirt; each layer has a group of tracks, each of which has the same order of like-counts/ play-counts and is connected to others in the same layer with edges whose weight is smaller than that of those linking the vertex to more central ones (=core). Therefore, thresholding edges based on weight might remove the edges within a layer but let edges connecting them to the inner layer or core remain.

6 Conclusion

In this study we have investigated the music network of SoundCloud, which was one-mode projected from the user-music bipartite network, whose data was sampled via SoundCloud API. The weighted version of the agglomerative algorithm found 9 communities with the modularity of 0.464. However, predominant communities, namely group 1,2 and 4 were analyzed exclusively. Based on the assumption that the genre of music group can be represented by the top 5 of the most frequent genre or tag information, group 1 and 4 belong to HipHop/R&B and group 2 belongs to House/Electro genre. The group 1 was found to have the highest like-counts likelihood among the dominant groups, which means music in the group could be liked more likely than those in other communities. The *core-periphery* structure was also observed within the network and could be attributed to the like-counts a track has obtained and thresholding edges based on weight of edges.

References

[1] R. Lambiotte and M. Ausloos, "Uncovering collective listening habits and music genres in bipartite networks," *Phys. Rev. E*, vol. 72, p. 066107, Dec 2005.

[2] A. Clauset, M. E. J. Newman, and C. Moore, "Finding community structure in very large networks," *Phys. Rev. E*, vol. 70, p. 066111, Dec 2004.

[3] M. E. J. Newman, "Analysis of weighted networks," *Phys. Rev. E*, vol. 70, p. 056131, Nov 2004.