Optimizing Local Record Labels

- via Analysis of Pittsburgh's SoundCloud Network Degeneracy -

Rohan Arepally
Akaash Mungale
Christoffer Thygesen

Introduction

From the various jazz and blues speakeasies to the rock and roll dive bars to Heinz Hall, home of the Pittsburgh Symphony Orchestra, music is one of the defining characteristics of Pittsburgh culture (1). Pittsburgh has recently emerged as an important hub of modern music as well. For example Pittsburgh rapper Wiz Khalifa's track "Black and Yellow", the colors of the city, reached number one on the Billboard Charts (2). We explore two trends in the evolving music scene that have risen since the development of the personal computer and the internet - electronic music and independent labels. According to Beatport, the world's largest music store for DJs, Pittsburgh is ranked the fourth in the "Next 13 Cities Ready to Rule Dance Music" (3).

With the advent of the internet, far more artists than ever before are finding success promoting their music through smaller independent labels. Therefore we believe that we can leverage the talent and untapped potential of the electronic musicians in the Pittsburgh area to create a highly successful new independent label.

Stage AE

Carnegie Music Hall

Three Rivers Arts Festival

Objective

In order to create the strongest Pittsburgh label possible, we want to select the optimal set of local artists to hire. Within the record label, we seek to maximize the interconnectedness of our artists. This is because the artists must be able to collaborate - by maximizing the overlap of each artist's following, we maximize the artists' overlap in music styles and familiarity with each other's works.

Initial Problem Statement

Create an optimal independent record label of 5 artists from the Pittsburgh Music Scene subject to a fixed minimum and maximum size

By solving such a problem, we can help independent record labels improve methods of choosing artists to hire to improve business. These results might prove useful for other players in the music industry.

Method of Solution

To best understand how artists in the area are interconnected, we want to find the intersection of independent musicians and internet based music promotion. To this end, we choose to pull data of Pittsburgh artists from SoundCloud - a popular social media website intended for musicians and music fans alike - using Python.

In order to model and analyze the seemingly ambiguous concept of the local indie music scene, we will use R to abstract the local SoundCloud network to an undirected graph, in which vertices represent Pittsburgh artists and edges imply the connected artists follow each other on SoundCloud. We will then attempt to apply k-core decomposition to further understand the centrality and connectedness of the graph and choose our artists according to the knapsack problem below.

Formulation: Artist Selection

$$\text{Maximize} \quad \sum_{j=1}^{n} k_j x_j$$

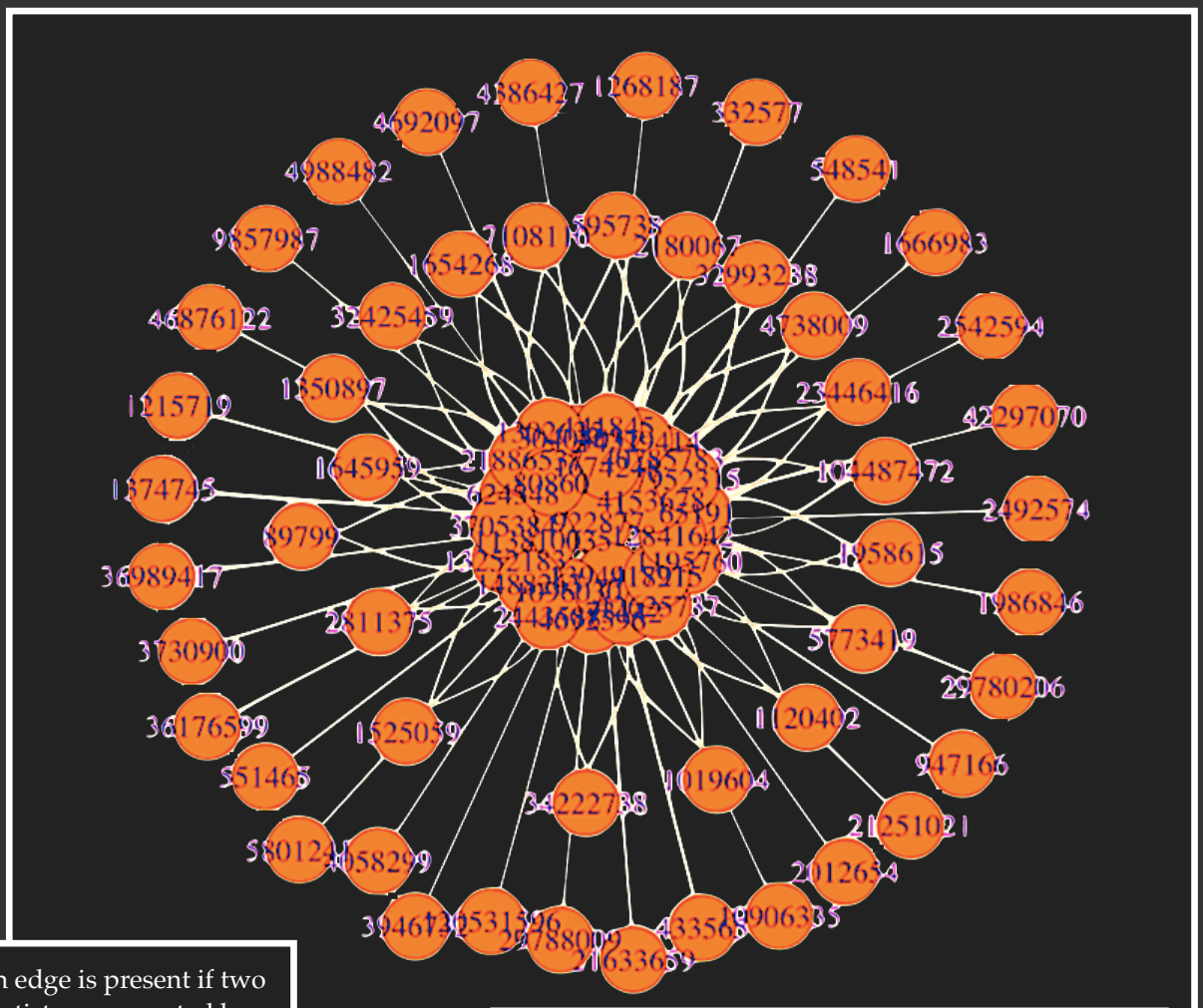
$$\text{Subject to} \quad x_j \in \{0, 1\} \ \forall j \in \{1, ..., n\}$$

$$S_{min} \leq \sum_{j=1}^{n} x_j \leq S_{max}$$

- n : candidate list size
- x_j : indicator variable for artist j
- k_j : k-coreness of artist j
- S_{min}, S_{max} : min, max label size

Data Collection and Structuring

In order to collect our data, we utilized the Soundcloud API (4) to search for artists that fit into our constraints. We first started with a seed pool of eight artists, chosen after interviewing VIA to better understand the Pittsburgh music scene. Using python, we organized the artists using a hash table where the key is the artist ID, issued by Soundcloud. The values stored in the hash table are structures consisting of information pertaining to the artists, such as artist name, location, number of followers, and followers.



An edge is present if two artists, represented by vertices, follow each other.

Visualization: Local Soundcloud Network

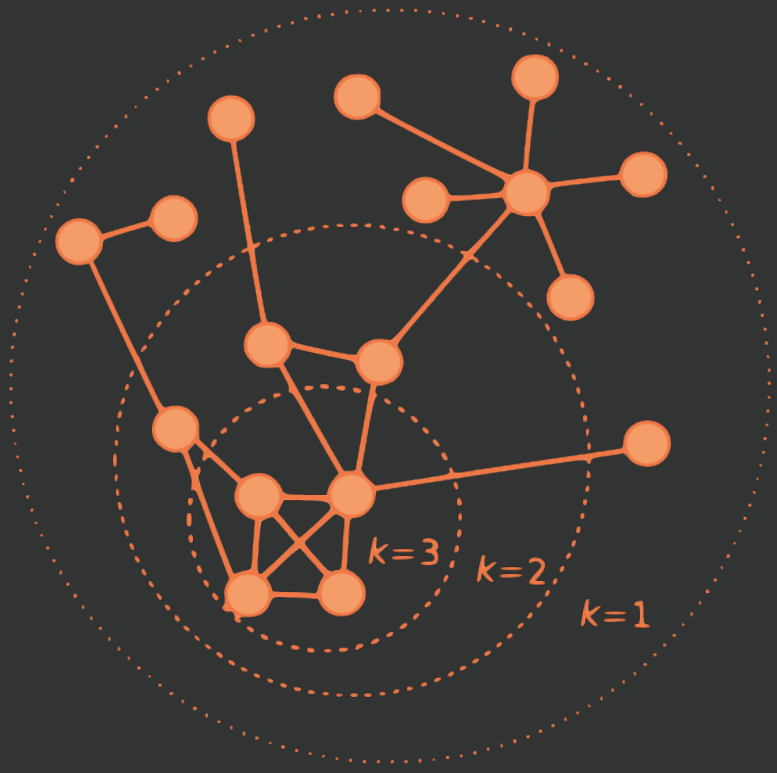
Using our initial seed, we did a breadth first search for two levels to gather data for our sample. We choose to iterate through only two levels due to the rate limiting imposed by Soundcloud, since we could query only a certain amount in a given time, we decided to start with a graph relatively small sample size, but large enough to be sufficient. As we indexed

the followers of a given artist, we filtered the followers who set their location to Pittsburgh. Furthermore, we made sure artists encountered in our breath first search had at least 100 followers so that we could filter out unused or spam profiles.

From this search we built an adjacent list to represent our graph, visualized in the figure above. An edge in our graph is present if two artists, who are represented by vertices, follow each other. We exported our data using the .csv format so we could conduct graph analysis in R.

k-core Decomposition

The k-core decomposition method, also referred to as the k-shell decomposition method, allows one to assess the hierarchy and cohesiveness of an undirected graph by producing a layered set of subgraphs. These subgraphs, called k-cores, are defined as the maximal connected subgraphs in which all vertices have degree at least k. k-cores are determined by ranking the vertices according to an integer k, such that k increases from the periphery of the network to the center.



Visualization: k-cores

The outermost dotted ring encompasses the network's 1-core, while the vertices between the two inner rings compose the 2-core. Those in central ring constitute the maximum k-core, in this case k = 3.

To achieve this decomposition, we start by iterating through our graph to find and remove all vertices with a degree = 1, and we assign them a rank of k = 1. This procedure repeats until the graph exclusively has vertices with degree $k \geq 2$ left. Subsequently, The procedure then recurses on the reduced graph and removes all vertices with degree = 2 and assign them a rank of k = 2. This routine is applied until all vertices in the graph have been ranked, and thus assigned to one of the k-cores. This rank is defined as the k-coreness: a vertex has k-coreness if it belongs to a k-core but not to any (k+1)-core. Note that k-cores are not necessarily connected. This decomposition process can be completed in polynomial time —a major reason for its utility in analyzing large networks.

Solution

Using R, we apply the k-core decomposition method and produce the k-coreness of each artist in our data set, shown in the output below:

```
> graph.coreness(g)
  922877   548541  4988482 46876122 13944182  1958615  4692596   595738 104487472
      27        1        1        1       27        2       26        2        2
32993238    73142  2180067  1645959  5773419  1120402 32425459  1019604 23446416
       2       26        2        3        2        2        2        2        2
34222738  1654268  1350897  1525059  4738009  2811375 30404031  1195760  1096030
       2        2        2        2        2        3       26       26       26
   65191   624348    89799  7108110  4692097  5801241  1268187 40785783 113810035
      26       26        2        2        1        1        1       25       25
13252183 42625737 14329414   917915 12841642   141845  4153678  1302625  1488263
      25       25       25       25       25       25       25       25       25
 1674248  7952315  2443588  3705384 21886537    80860  3730900 36989417  1666983
      25       25       25       25       25       25        1        1        1
  433565 29780206  4058299  9857987 10906335 21633659  2542594 42297070  2012654
       1        1        1        1        1        1        1        1        1
36176599 29788009 120531596  3946732  4386427 21251021   947166  1986846  1215719
       1        1        1        1        1        1        1        1        1
 1374745   551465  2492574   332577
       1        1        1        1
```

R Output Format

- ID_j : Soundcloud ID of artist j
- k_J : k-coreness of artist j

To select our artists, we hone in on the artists in our set with maximal k-coreness. Given the nature of our sample size and our two constraints - that each artist can either be included

or excluded and that our label size should be at most 5 - we can just apply the greedy algorithm. We will use this approach with discretion.

We compare all artists with maximal k-coreness with one another: Mike Masai, VIA, Buku, C. Scott, sub Mistress, Edgar Um, dwayne rifle, Troxum. We remove VIA from our list since they are already a Pittsburgh collective. We also remove Buku since he is disproportionately more successful than all the other artists. Lastly, we remove C. Scott whose style of music, hip hop, deviates from the other electronic artists. This leaves us with our finalized collective, shown in the table below:

Artist	k-coreness	# Followers	# Following
Mike Masai	26	511	929
sub Mistress	26	1105	1082
dwayne rifle	26	420	594
Troxum	26	226	1859
Edgar Um	26	418	1088

Conclusion and Discussion

In summary, we constructed a network of artists and defined and found a method to connect the artists. Finally, we analyzed this network by solving the problem of creating an independent internet based artist collective, a relevant and increasing popular option for new artists (5). We then solved our initial problem and found a collective of Pittsburgh based artists, by maximizing their connections to the community of local musicians and each other. We even discovered that our analysis could prove useful to more of music industry than previous intended: not just record labels looking for talent, but venues looking for musical acts that synergies best to put on a show bill together, and artists interested in knowing which major cities their followers lived in, so that they could focus promoting and performing efforts in those locations.

Our approach of modeling a collection of local artists as a network and applying k-core decomposition was effective but not perfect - our algorithm for selecting the optimal subset of local artists could still use improvement due to a problems such as a small sample size, selection bias in data collection, and lack of information and computing power.

We can see from the construction of the graph that our graph is rather overly connected - the middle cluster of artists is primarily comprised of the most connected seed artists and the artists found by the first iteration of breadth first search. Our data selection method of interviewing VIA to find our seed artists resulted in a bias towards the electronic music scene in Pittsburgh. Further studies with additional computing power would use a larger, more random selection of seed artists across a more diverse set of genres for a more significant sample size.

Furthermore, while k-core decomposition is still quite powerful for analyzing the centrality and hierarchy social networks, it only works for undirected graphs. In the process of abstracting the local SoundCloud network to an undirected graph, we made the critical assumption that if there is a mutual following between artists, they are either interested ion n working together or that their musicals styles share some degree of overlap. This assumption simplified the computational complexity and analysis process, but prevented any analysis of one-way follows between artists.

Lastly, it is important to note that proper models of social networks like SoundCloud are weighted, as not every connection is created equal, in reality; one could run algorithms not restricted to polynomial runtime to get more detailed analysis of our data an algorithm, such like a weighted k-core decomposition method would allow one to make a more realistic model and thus draw more significant conclusions.

1. https://sites.google.com/site/pittsburghmusichistory/
2. http://www.billboard.com/articles/news/473134/wiz-khalifas-black-and-yellow-tops-hot-100
3. (https://news.beatport.com/us/the-next-13-cities-ready-to-rule-dance-music/)
4. https://developers.soundcloud.com/docs/api/reference
5. http://www.telegraph.co.uk/culture/music/9672807/The-record-label-is-dead-long-live-the-record-label.html
6. (http://via-hq.com/)