BRIDGING DATA AND ELECTRONIC DANCE MUSIC

by

Jasdev Singh

Submitted in partial fulfillment of the
requirements for the degree of
Bachelor of Science Computer Science

at

University of Virginia
Charlottesville, VA
March 2014

UNIVERSITY OF VIRGINIA

DEPARTMENT OF COMPUTER SCIENCE

The undersigned hereby certify that they have read and recommend to the Faculty of Undergraduate Studies for acceptance a thesis entitled "BRIDGING DATA AND ELECTRONIC DANCE MUSIC" by Jasdev Singh in partial fulfillment of the requirements for the degree of Bachelor of Science Computer Science.

Dated: March 9, 2014

Supervisor: _____
Dr. Yanjun Qi

Reader: _____
Matthew N. Eisler, PhD

UNIVERSITY OF VIRGINIA

DATE: March 9, 2014

AUTHOR: Jasdev Singh

TITLE: BRIDGING DATA AND ELECTRONIC DANCE MUSIC

DEPARTMENT OR SCHOOL: Department of Computer Science

DEGREE: B.S. CONVOCATION: May YEAR: 2014

_____

Signature of Author

Table of Contents

Abstract

Electronic dance music (EDM) is a relatively new genre with unique features such as a large number of subgenres and the notion of playing live events in a festival format. With its rise to mainstream popularity, there are opportunities to explore the social dynamics of EDM by analyzing datasets associated with the genre. In our research, we seek to retrieve information from three such datasets, the DJ Magazine Top 100, live festival set lists, and SoundCloud comment data. The data mining techniques presented in this thesis provide great potential for helping listeners, musicians, and online storefronts, e.g. automatic recommendation and music summarization.

Acknowledgements

- Dr. Yanjun Qi, PhD, Department of Computer Science
- University of Virginia Association of Computing Machinery

Chapter 1

Introduction

1.1 Background on dance music

Electronic dance music (often called EDM) includes multiple electronic music genres, usually for dance-based events (festivals and clubs). At these live events, it is very common for disk jockeys (DJs) to play continuous sets, ranging from 30 minute to multiple hour mixes. Moreover, producers and DJs will often collaborate on and play each other's music. Below is a graph from a recent report by Google Music's Research Group[1] showing the popularity of various genres uploaded to the Google Play store[2], since the service's inception. Highlighted by the red box is Dance and Electronic music. It can be seen that EDM is roughly the 6th largest genre, as of the most recent year in the survey.

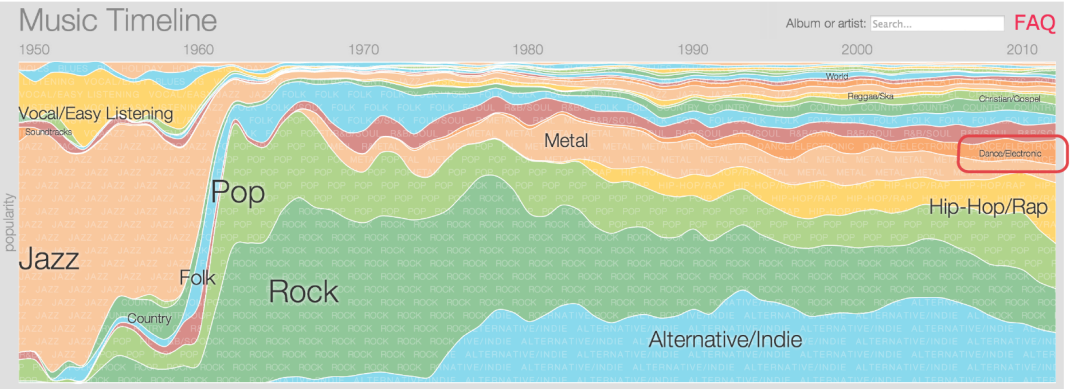

Figure 1.1: Google Music genre timeline - rise of dance music

A unique feature of dance music is the wide array of subgenres present. This allows for the rise of communities of producers and opportunities for remixing originals into other subgenres. Zooming in on the graph presented previously, we can see the diversity of the overall genre, with notable examples being trance, techno, house, ambient, and drum & bass.

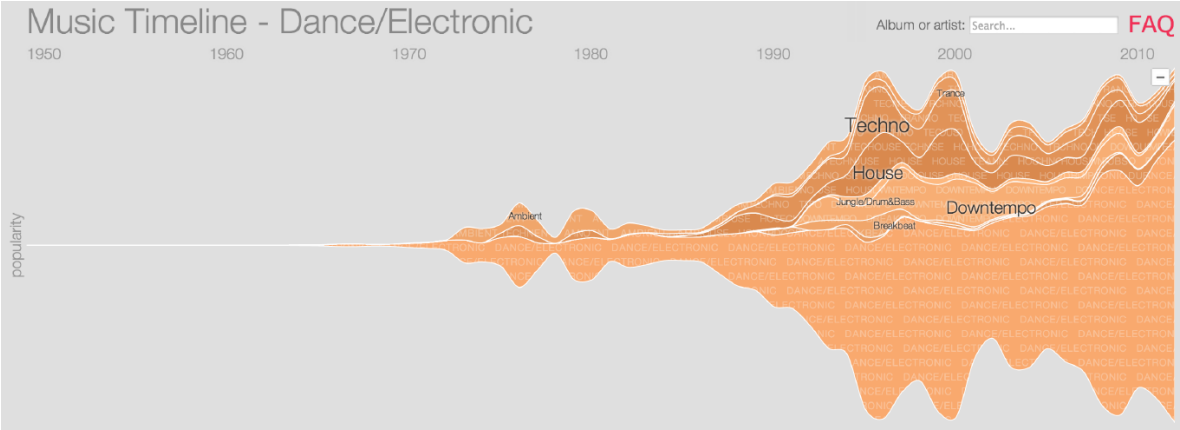

Figure 1.2: Subgenres present within Dance music

Next up, we will highlight some of the most popular genres within EDM. To begin, House is a genre typically identified by a 4 by 4 kick drum and a 120 - 130 beats per minute (BPM) range. Some common names within House are Avicii, Swedish House Mafia, and David Guetta. While bucketing artists into genres is not an exact science, one can generally associate artists with genres based on the majority of their productions. Secondly, Electro is a style with identifiable synths and a tempo range of 125 - 135 BPM. Some notable artists in the space are Dada Life, Justice, and Boy's Noize. Third is Dubstep, defined by an oscillating wobble bass. The BPM range hovers around 140 and notable producers in the genre are Skrillex, Bassnectar, Skream, and Benga. Lastly, Trance is another large subgenre popular around the globe. Its tempo range is wide, 125 - 150 bpm, and is noted by anthems and exaggerated builds. Trance has been doing extremely well in the DJ Magazine Top 100, which we will introduce soon, and is lead by artists such as Armin van Buuren, Above & Beyond, and Ferry Corsten.

1.2 Motivation for study

Being a relatively young genre, there has not been a lot of research done on the EDM social demographic and artists. In this paper, we will take a look at three datasets to gain information about this style of music. To start, we will note trends in the DJ Magazine Top 100 poll, which is an attempt to put a stack rank on DJs and producers from year to year. Insights on this chart can help us to better understand

dynamics such as the rise and fall of genres and record labels. Secondly, we will look at set lists of what DJs played at live events, specifically one festival in New York called Electric Zoo. From these set lists, we can make connections from specific DJs to others by how they played and/or remixed each other's work. Such connections naturally lead to the formation of a network diagram. From such diagrams, we can perform an analysis on communities present to gain insight on the underlying social constructs. Lastly, we will perform a sentiment analysis on public user comments on music hosted on the popular online streaming service, SoundCloud. From the text data housed in these timestamped comments, we will attempt to form a heuristic for the overall positivity towards the underlying audio at that portion of the music. This is hugely beneficial for two major reasons. First, given any song, we can now attempt to answer the question of 'what is the best portion of this song?' via the text comments as a proxy for the audio. Extending this, an answer to such a question can help in summarizing larger pieces of audio or even serve as the recommended preview, prior to buying music on online storefronts. To recap, our motivating factors are as follows:

- Better understand the social dynamics between entities in the genre
- Retrieve information about musical quality via sentiment analysis
- Construct a baseline 'best window' algorithm to be used by online storefronts

This research will help better inform artists and musicians about how their work fits in the larger network of Electronic Dance Music as a whole. Additionally, listeners and online music storefronts can benefit from our 'best window' algorithm developed in this paper to preview tracks and better present content to users, respectively.

Chapter 2

Previous work

2.1 Research in EDM

Previous work in the realm of analyzing data on EDM is far and in between. However, Eventbrite, an online ticketing platform for live events, put out a study in June 2013 showcasing behavioral trends on how EDM listeners shared content on social media before and after events[3]. This serves as a perfect contrast to our study of how DJs play music at live events, giving us a holistic view. Some of the key takeaways from the study were:

- EDM fans are twice as likely as other music fans to want to attend an event when their friends post on social media
- Over half of EDM fans revealed that they would pay for live streaming of an EDM concert
- People who attend EDM events are significantly more likely to share on social media before (67%), during (41%) and after (63%) an EDM event than people who attend music events but not EDM events (with 37%, 21%, 51% respectively)

2.2 Research in sentiment analysis

Sentiment analysis is a branch of Natural Language Processing that we will rely heavily upon in our attempt to gain insights from text data present on the music streaming service, SoundCloud[4]. More specifically, sentiment analysis attempts to determine the attitude and polarity of a document in both supervised and unsupervised contexts. In this paper, we will employ a supervised machine learning technique with a training corpus of thousands of SoundCloud comments to detect positive and negative emotions. As a field of study, sentiment analysis has gained popularity with

the rise of social media, as the amount of reviews, ratings, and recommendations have proliferated.

Chapter 3

Overview of data sources

Figure 3.1: Three datasets in which we will perform data mining - DJ Magazine Top 100 (left), 1001Tracklists (middle), SoundCloud (right)

3.1 Dataset 1 - DJ Magazine Top 100

Using data available within the community, we seek to convert it into knowledge using a variety of analysis techniques. To begin, we are going to use poll data from the DJ Magazine Top 100. Starting in 1997, DJ Magazine has put together a popular vote poll to determine the top 100 DJs each year. However, when it comes to subjective topics, like music, this type of poll has received criticism since popularity does not always coincide with musical talent. But, on a higher level, we can use this time series data to draw larger and more general conclusions about the genre and its constituents. In our study, we scraped the poll data and made it available in both CSV[5] and JSON[6] format. Some of the challenges faced when scraping this data include:

- Resolving accented characters in DJ names
- Naming consistency resolution (e.g. 'DJ Snake' versus 'Snake')

- Artist stage name changes (defaulted to latest known stage name)

3.2 Dataset 2 - 1001Tracklists

Second, we will be looking for connections between DJ set lists that are available on the popular site, 1001 Tracklists. As mentioned before, it is common for DJs to play longer sets during live events. This site serves as an aggregator for the tracks played in such sets to better allow for song discovery. Utilizing this, we will scrape the tracklist data for a specific event, Electric Zoo NYC[7], and try to establish connections between artists by the music they play in each others' sets. This can allow for the creation of a community diagram, where we can logically see higher level connections between groups of DJs. In parsing this data, JSON was used in a manner reminiscent of an adjacency list, typically used in graphical settings. See below for a sample entry:

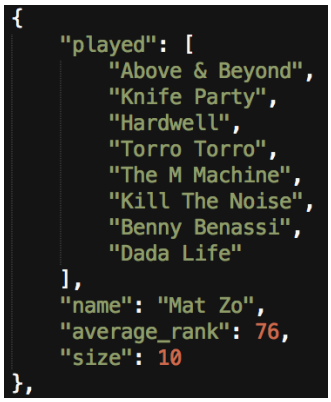

Figure 3.2: Sample schema for describing social networks among DJs

In the figure above, the artists in question is Mat Zo, as indicated by the *name* key. Ranking information such as the *average_rank* and *size* are used to scale the nodes in the diagram. Lastly, the *played* key maps to an array of other unique artist names, whose work appeared in Mat Zo's set at the festival. Using this schema, we can generate a network diagram of artists at an event, to attempt to gain information about underlying group dynamics.

3.3 Dataset 3 - SoundCloud comments

Lastly, we will mine text data from the music streaming service SoundCloud. The motivation for this inquiry is the fact that comments on the service are timestamped at specific points during songs. So, using sentiment analysis, we can seek to get a better understanding about the underlying audio of a track, given what people are saying in text. Our initial data mining fetched over 200,000 of these comments and thousands of them were labeled manually as being either negative, neutral, semi-positive, or really positive in emotion. These labeled points were then used to train a classifier to distinguish sentiment on new unlabeled comments.

Chapter 4

Social Dynamic Analysis of DJ Magazine Top 100

4.1 Top 100 Case Study #1, Swedish House Mafia

To exemplify some of the trends in these charts, we begin with the DJ trio that is Swedish House Mafia. Starting in 2008, Steve Angello, Sebastian Ingrosso, and Axwell began touring around the world together and recently stopped in 2013. Below is a plot of their respective rankings in the chart, including the ranking of the group overall:

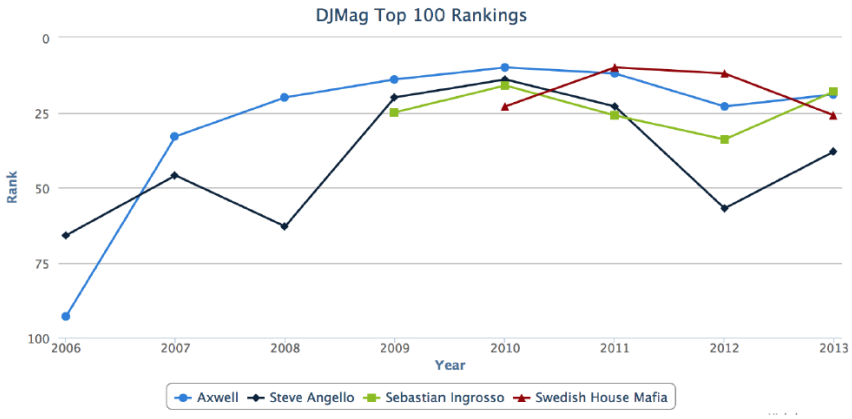

Figure 4.1: Temporal rankings of Swedish House Mafia Members

From this figure, we can see that upon the group' splitting in 2013, both Sebastian Ingrosso and Axwell had higher rankings in the poll, even above the group, whereas Steve Angello's regard in the public eye was ranked lower. The reasoning for the group's spit is unknown (although a documentary `http://leavetheworldbehind.com/` is set to be released soon), but this may help us gain some insight on the situation.

4.2 Top 100 Case Study #2, Trance rankings

Trance is an extremely popular electronic genre that has typically held a strong showing in the polls. However, in 2013, there has been some pushback from fans against the genre, as DJs and producers have been experimenting with more "radio" style tracks. Below we have a plot of all the trance DJs appearing in the Top 100:

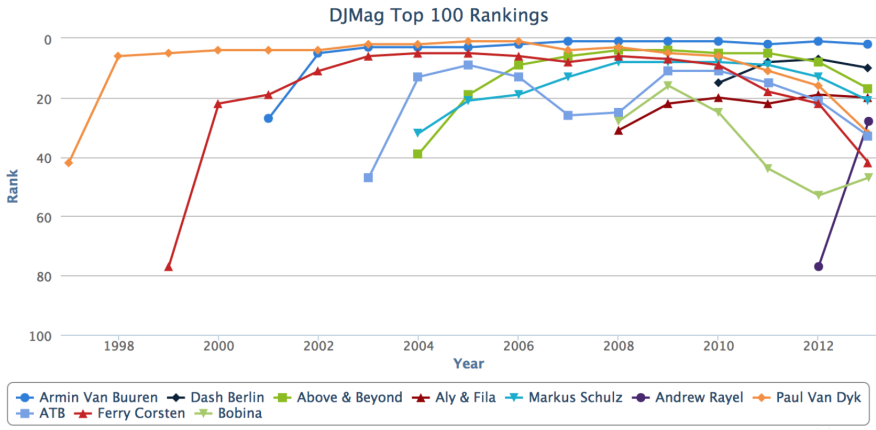

Figure 4.2: Temporal rankings of all DJs related to the Trance genre

In 2013, every Trance DJ in the poll, with the exception of Bobina and Andrew Rayel, actually declined in ranking, when compared to the previous years.

4.3 Top 100 Case Study #3, Mentor rankings

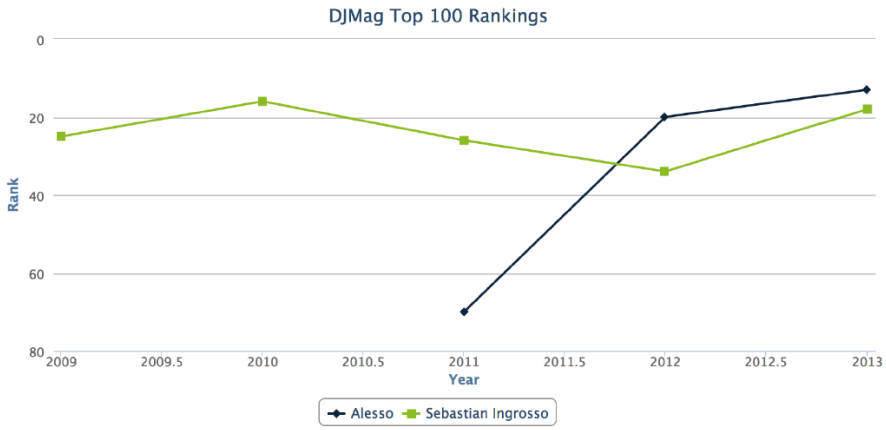

Figure 4.3: Temporal rankings dynamics of a pair of mentor/mentee DJs

Alesso, a 22 year-old DJ from Sweden, got an early start in 2010. Later on in his career, he received mentoring from the aforementioned Sebastian Ingrosso. This mentor / mentee relationship has been interestingly reflected in the Top 100 charts. From the graph below, you can see that Alesso's rank actually surpassed Ingrosso's, despite him being the mentee.

4.4 Top 100 Case Study #4, Record labels

Like other genres, it is common in EDM for artists to come together in groups to form record labels. Applying this grouping to our Top 100 data, we can see the performance of various labels in the poll.

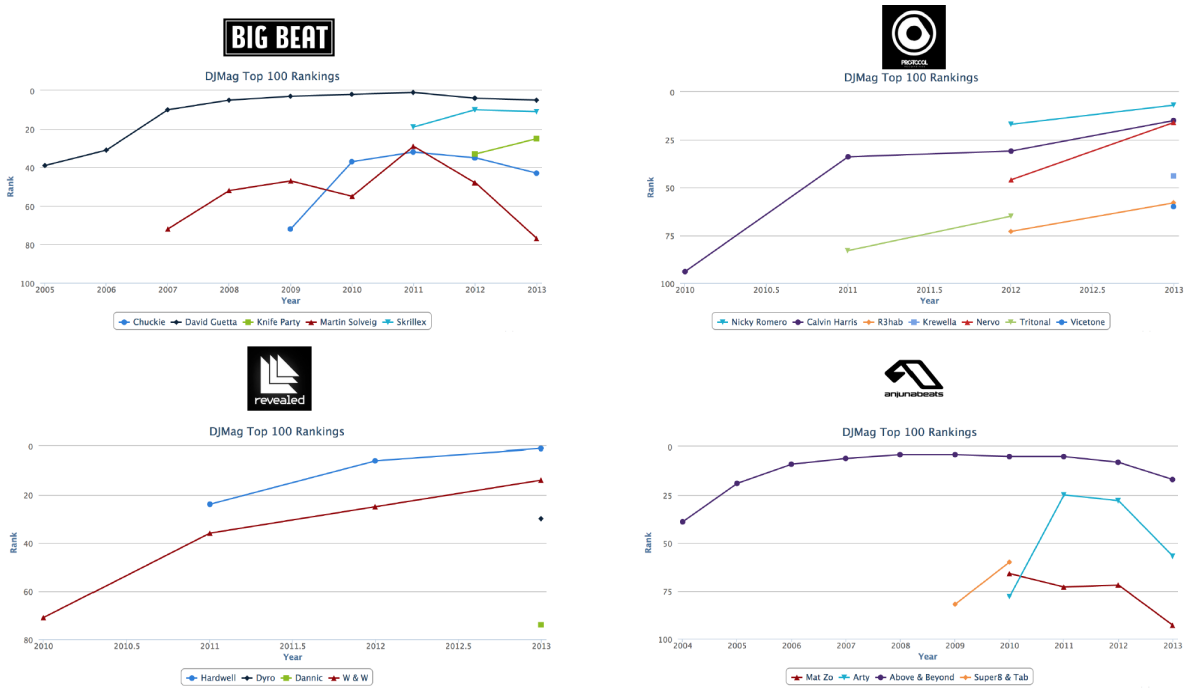

Figure 4.4: Temporal rankings dynamics of record labels

These graphs can help us see how the label is performing overall in the charts and identify integral artists.

Chapter 5

Social network analysis of festivals and live events

Large scale EDM events have become increasingly common, with upwards of 300 thousand attendees in some cases. During such events, DJs will often play sets composed of not only their work but work from other artists as well. This scenario lends itself well to a sort of network analysis. To get tracklist data, there is a crowd-sourced site, 1001Tracklists that hosts set lists from a wide array of events and podcasts. Moreover, users can go in and attach audio samples to tracks, link unknown songs to other appearances in different contexts, and even resolve disputes regarding an unknown song via voting. To use this data, we scraped the site into JSON and crossed it with Top 100 rankings to get a bigger picture of community dynamics within festivals. Attempting to showcase this, we decided to use data from an annual festival run in New York, Electric Zoo. Below is a graph of artists present in the festival, with the font of their names scaled by their respective Top 100 ranking. Moreover, a black colored label is the current source DJ, with green edges representing the origin DJ playing a song by the source DJ and the red edges represent that the source DJ played a song by the destination DJ. An interactive version of this diagram is available online[8].

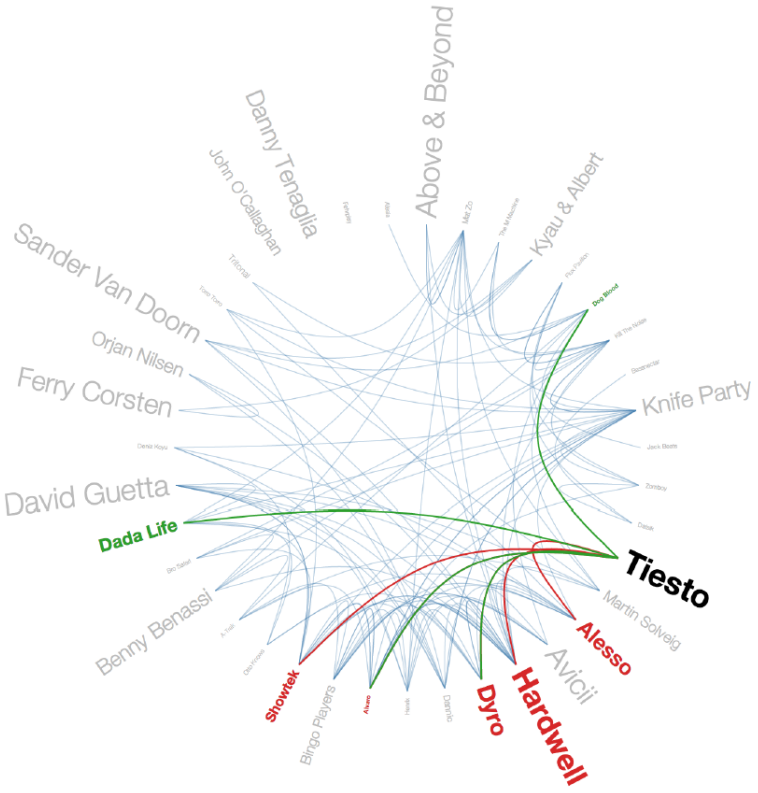

Figure 5.1: Social network diagram of DJs at a festival event called Electric Zoo

5.1 Electric Zoo Highlights, opportunities for breakout DJs

A great opportunity for breakout artists at these events is the scenario where a highly ranked DJ plays one to their songs. This is desired because headliners usually have slots on bigger stages, when compared to younger artists, who typically perform on side stages with a smaller audience. To highlight such a scenario during Electric Zoo, Avicii played a song by the young duo Dog Blood.

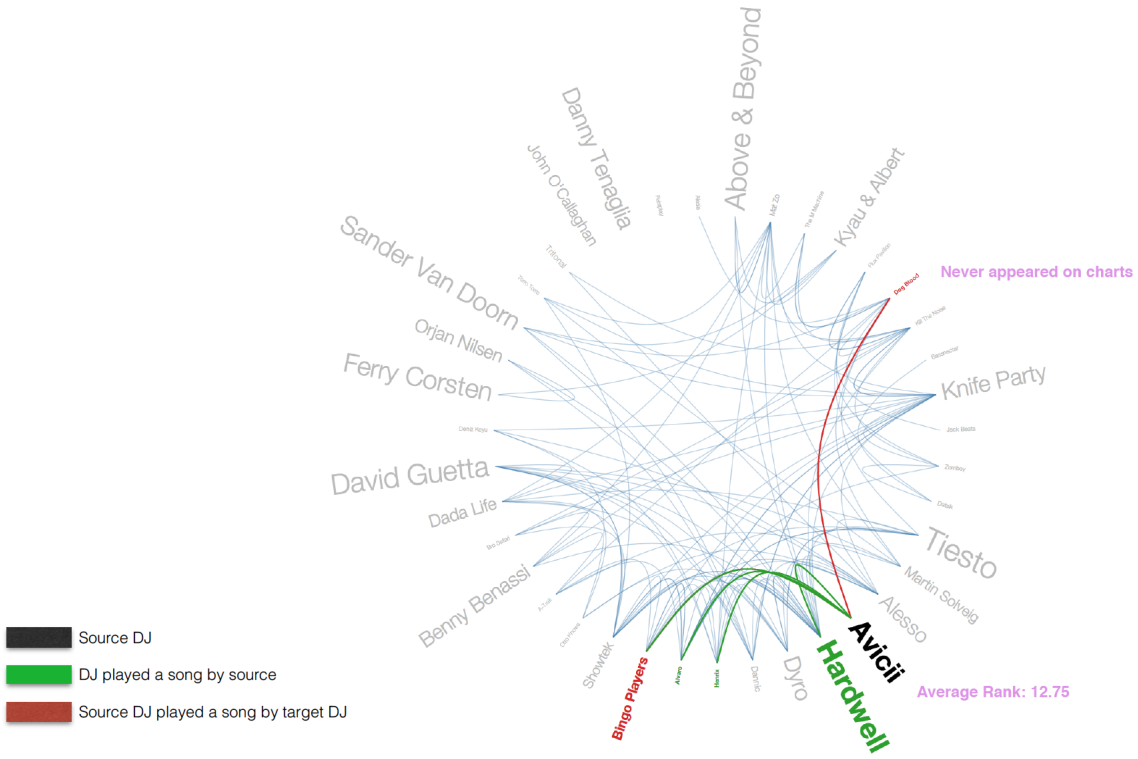

Figure 5.2: Example of a breakout artist (Dog Blood) being played by an established DJ (Avicii)

As noted in the diagram, Avicii's average rank in the charts is 12.75, whereas Dog Blood has never actually appeared on the poll. This edge marks a huge opportunity for the duo to gain an audience and following.

5.2 Electric Zoo Highlights, Sets with Isolated Tracks

Moreover, with this diagram, we can notice specific DJs who are disjoint in connections to any other DJs present in the festival. Below, John O'Callaghan, Danny Tenagila, and Fehrplay all had set lists that were disjoint from others in the festival. This is likely attributed to their genres being relatively niche.

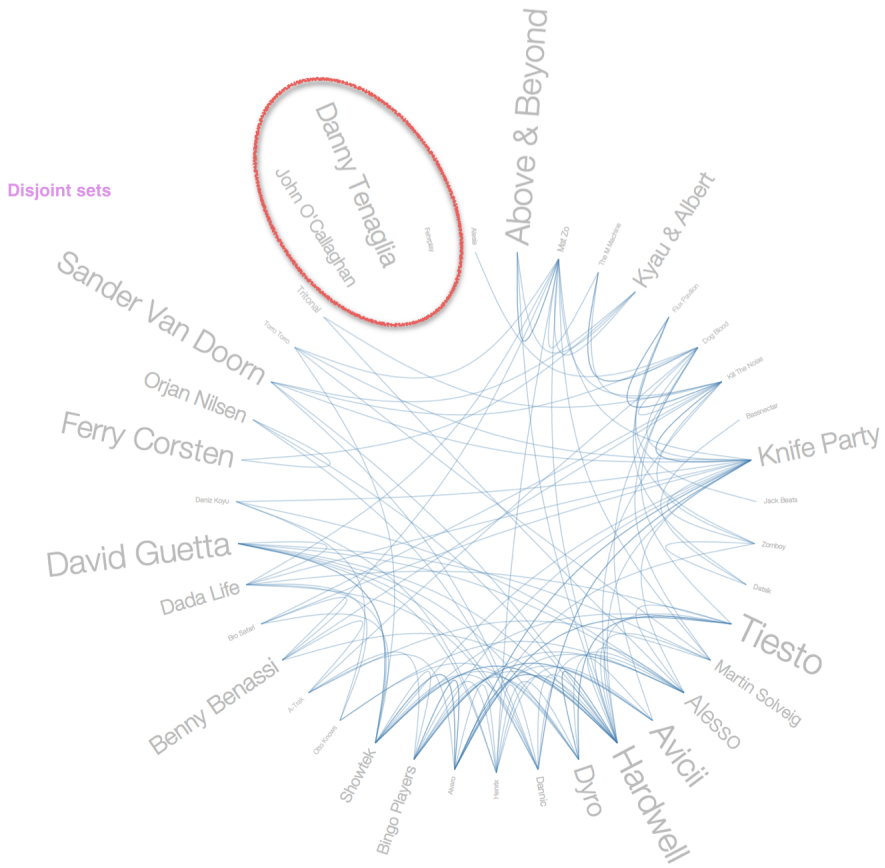

Figure 5.3: Social Network Diagram showing artists with disjoint sets in a live event

5.3 Electric Zoo Highlights, most played artists

Lastly, we can seek to answer the question of which artist was most played at the festival. This is calculated by iterating over all artists and keeping track of the one with the largest number of incoming edges. From Electric Zoo, Knife Party had the most appearances in other sets with 12 plays.

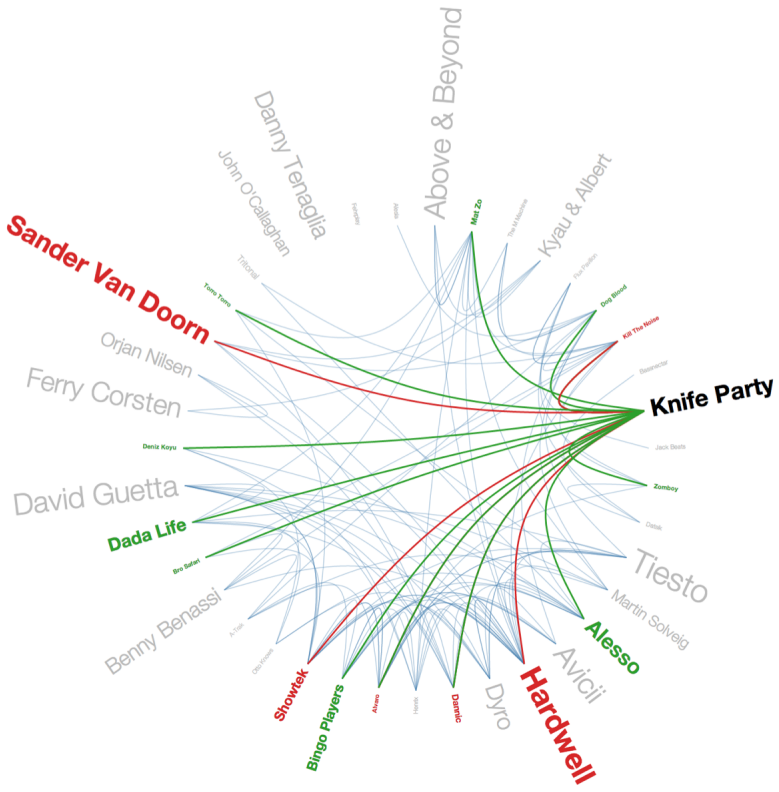

Figure 5.4: Social network diagram showing the most played DJ (Knife Party) at a live event

Chapter 6

SoundCloud Sentiment Analysis

6.1 Data mining overview

To begin training an accurate classifier in determining sentiment, we need a lot of training data. Using SoundCloud's public REST API[9] we wrote a script to fetch 200,00 comments on songs in the EDM genre. These were stored locally in flat text files for labeling.

6.2 Training process

Prior to manually labeling the comments collected, a few steps were taken to clean the data. First, an unfortunate part of the SoundCloud service is the presence of spammers who tend to comment with links to his/her music, to gain views. Such data would interfere with our core sentiment analysis, so all comments with URLs were stripped from parsing. Additionally, non-English comments were removed to specifically focus on deriving sentiment in the English context. Lastly, all text was converted to lowercase and stop words (words that do not add sentiment value such as 'the', 'is', etc.) were removed. A special note about the stop words is that both a standard list of English stop words and a custom stop word list were used. The English stop word list was taken from the default module provided in the Python Natural Language Toolkit (NLTK). The need for an additional custom stop word list is because in EDM there are common words, used in context, that do not add sentimental value. Examples of these words are 'remix', 'listening', and 'tune.'

After this cleansing process, the comments were then manually labeled by splitting them across four text files, representing negative, neutral, semi-positive, and really-positive comments. The logic behind this range scale as opposed to a binary scale is that sentiment often falls along a gradient as opposed to being clearly positive or

negative. Splitting the comments up in separate text files allowed us to load them into memory later, labeled by their respective filename.

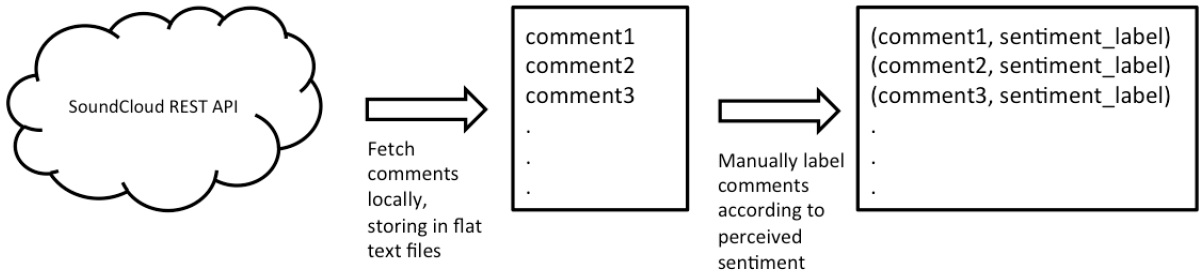

Figure 6.1: System diagram of how SoundCloud comment data was fetched and classified. Step 1: Given a list of hundreds of SoundCloud URLs, fetch their respective comments and save locally in a flat text file. Step 2: Manually label each comment along the scale of negative, neutral, semi-positive, or really-positive according to sentiment

6.3 Tools used

To perform our sentiment analysis, we used a Naive Bayes Classifier from Python's Natural Language Toolkit[10]. This classifier takes in our labeled comments from above and creates a frequency table of common tokens and their appearance in specific labels. For example, the token 'great', appeared in many really-positive comments, so when a new comment is introduced to the classifier, the presence of the word 'great' will likely bump the comments label towards being really-positive. Below is a sample of some of the informative features of the classifier:

Token	**Label Ratio**	**Value**
bad	negative : semi_positive	15 : 1
holy	really_positive : semi_positive	15.1 : 1
beautiful	really_positive : semi_positive	11.2 : 1
best	really_positive : semi_positive	7.9 : 1
nice	semi_positive : really_positive	6.8 : 1
stop	neutral : semi_positive	4.3 : 1

Table 6.1: Top 6 informative features of the Bayes Classifier and the associated label ratio values

6.4 Best window algorithm

With our trained classifier at hand, we can begin our analysis of SoundCloud songs. Below is a diagram of how comments typically show up during playback:

Figure 6.2: How text comments are rendered on SoundCloud tracks

Using these comments, we seek to answer the question of what window of this song is the 'best' in terms of user sentiment, as derived from the text comments. We can accomplish this in the following steps:

1. Given a SoundCloud URL, use the REST API to fetch all timestmped comments associated with the track

2. If we seek to find the best 10 seconds, we create a sliding 10 second window across the duration of the track

   (a) For each 10 second window, we score each comment in the window according to its classified label outputted from the Bayes Classifier. Attributing a score to each label (e.g. negative is -1, neutral is 0, semi-positive is +1, and really positive is +2). Summing these weights we get a score for the entire window.

3. Keep track of the window with the highest score from above loop and output as 'best' portion of track.

6.5 Visualizations and extensions

Running the algorithm, we can create a heat map of how comment sentiment is distributed along a given track.

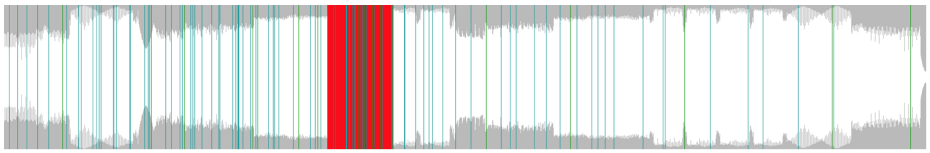

Figure 6.3: Heat map of sentiment of SoundCloud comments. Red region indicated the portion of the track with the highest overall sentiment score

Above, the lines represent user comments and the closer the comment is to the color green, the more positive it is in sentiment. Additionally, the window in red represents the best 8 second window of the song via the algorithm above. To extend this notion of extracting the best parts of songs, we took 10 of the top 100 songs on Beatport's Top 100[11] and ran this algorithm on them. From these 10 best portions, we created a single track composed of those best parts[12]. Lastly, from the text data, we can also visualize the comments in for form of a word cloud, where the font of a token is scaled by frequency.

Figure 6.4: Word cloud of tokens present in comment text data. Font is scaled by frequency of token in comment corpus.

6.6 Results

We report the Precision/Recall score to evaluate our experiments.

Precision: Among the negative (neutral or semi-positive or positive) comments detected by the system, how many are true negative comments ?

Recall: For all possible negatives (neutral or semi-positive or really-positive), how many were detected by system?

A good detector should have both high precision and high recall. If we let d be the number of negative (neutral or semi-positive or positive) detected, z be the number of negatives manually judged, and dz be the number of negative manually judged as negative among the detected ones, the precision and recall for the system are:

$$precision = \frac{dz}{d} \tag{6.1}$$

$$recall = \frac{dz}{z} \tag{6.2}$$

Comment corpus breakdown

	Negative	**Neutral**	**Semi-positive**	**Positive**
Total Count:	133	114	493	425
Percentage of Corpus:	11.4	9.8	42.3	36.5
Precision Rate:	1.0	1.0	0.601	0.932
Recall Rate:	0.083	0.088	0.982	0.744

Error Rates - Cross validation

	2-Fold	**5-Fold**	**10-Fold**
Error Rate:	0.146	0.167	0.423

Chapter 7

Conclusion and Future Work

From the three datasets analyzed above, we can better explore the trends and connections between the main actors within the EDM industry. From the DJ Magazine Top 100, the temporal nature of the popularity of genres is revealed as the rise and fall in the charts. Additionally, there are dynamics between artists, as set lists from Electric Zoo demonstrate that there exist communities of DJs who play each others music. Lastly, from analyzing text data from SoundCloud, we can reach a heuristic for the 'best' portion of a song via sentiment analysis on timestamped data points.

The last section, where we developed the 'best window' algorithm, as a heuristic to determine the portion of a track with the best overall sentiment score, is extremely valuable for both listeners and online music storefronts. Listeners can use this to better browse musical libraries by listening to trimmed previews of longer tracks via this algorithm. This has implications in content discovery, as users can now find music in a faster manner, by listening to key parts of music. On the online storefront side, companies can use sentiment data to effectively make better previews for content that are presented to customers before purchase.

Being a relatively young genre, there are many applications and extensions to the research presented. For instance, the best window algorithm can be used in contexts such as summarization of audio. Moreover, when online storefronts present previews of tracks, they can use the best windows as a default choice for clips to show users prior to purchase.

On a final note, all of the code and work from this project is open source[13].

Chapter 8

References

[1] http://research.google.com/bigpicture/music/

[2] https://play.google.com/store/music

[3] http://www.eventbrite.com/pressreleases/edm-fan-behavior-differs-greatly-from-that-of-other-music-fans/

[4] https://soundcloud.com/

[5] https://github.com/Jasdev/soundcloud-sentiment/blob/master/heatmap-gen/static/datasets/dj-mag-top-100.csv

[6] https://github.com/Jasdev/soundcloud-sentiment/blob/master/heatmap-gen/static/datasets/dj-mag-top-100.json

[7] http://www.madeevent.com/ElectricZoo/

[8] http://www.jasdev.me/thesis/ezoo

[9] http://developers.soundcloud.com/

[10] http://www.nltk.org/

[11] http://www.beatport.com/top-100

[12] https://soundcloud.com/jasdev-singh/beatport-top-10-drops

[13] https://github.com/Jasdev/soundcloud-sentiment