Modeling Musical Influence with Topic Models

Uri Shalit URI.SHALIT@MAIL.HUJI.AC.IL
ICNC-ELSC & Computer Science Department, The Hebrew University of Jerusalem, 91904 Jerusalem Israel

Daphna Weinshall DAPHNA@CS.HUJI.AC.IL
Computer Science Department, The Hebrew University of Jerusalem, 91904 Jerusalem Israel

Gal Chechik GAL.CHECHIK@BIU.AC.IL
The Gonda Brain Research Center, Bar Ilan University, 52900 Ramat-Gan, Israel

Abstract

The role of musical influence has long been debated by scholars and critics in the humanities, but never in a data-driven way. In this work we approach the question of influence by applying topic-modeling tools (Blei & Lafferty, 2006; Gerrish & Blei, 2010) to a dataset of 24941 songs by 9222 artists, from the years 1922 to 2010. We find the models to be significantly correlated with a human-curated influence measure, and to clearly outperform a baseline method. Further using the learned model to study properties of influence, we find that musical influence and musical innovation are not monotonically correlated. However, we do find that the most influential songs were more innovative during two time periods: the early 1970's and the mid 1990's.

1. Introduction

In the past few years, significant research has been invested in learning to organize and classify music, with the goal of allowing users to retrieve and discover new music (Turnbull et al., 2009; Su et al., 2010; McFee et al., 2012). This includes classifying songs into musical genres (Scaringella et al., 2006), creating playlists and recommending music (Logan, 2002; 2004; McFee & Lanckriet, 2011) and even searching music by humming or notes (Lu et al., 2001). The growing availability of music tracks and of methods to capture their unique acoustic signatures now opens new possibilities to study the structure of the music collective itself. Specifically, this paper provides a quantitative modeling approach to study musical influence. Musical influence is often discussed, but has never been studied quantitatively and at a large scale before.

Proceedings of the 30th International Conference on Machine Learning, Atlanta, Georgia, USA, 2013. JMLR: W&CP volume 28.

Although central to understanding musical creation, the concept of musical influence is loosely defined, and its role debated among scholars of art history and cultural critics. For instance Bloom claimed that important artistic work results when an artist creates original work against existing influence ("The Anxiety of Influence", Bloom, 1997), while Lethem claimed that "originality and appropriations are as one" in all artistic endeavor, ("The Ecstasy of Influence", Lethem, 2007). In a cultural-artistic landscape that is very much shaped by sampling, remixing, and copy-pasting, the question of the role of influence in art is always present (Reynolds, 2011). Unfortunately these questions were never studied in a data-driven way.

The challenge in modeling a whole musical corpus is two-fold: The audio signal itself is a complex continuous signal, with meaningful structure on multiple timescales; and there exist intricate and evolving relations between artists, songs, and genres.

By learning probabilistic models of the influence that a song has on later songs, this paper offers a quantitative measure of influence, that can be used to engage the ongoing discussion about influence with scientific and data-driven arguments. We use the learned models to detect influential songs automatically, and study the relation between influence and innovation.

Our model of music influence is based on *Dynamic Topic Model* (DTM) (Blei & Lafferty, 2006) and *Document Influence Model* (DIM) (Gerrish & Blei, 2010). These models were originally developed in the context of analyzing text documents and used to analyze how the language of scientific papers evolves. Under the DIM, an influential scientific paper is one whose language is adopted by its successors in its scientific field. In our case, of music influence, the audio content of songs replaces the text of a scientific paper, and we consider a song to be influential if its “musical-language” (or sound-content) has been adopted by later songs in related genres.

We find that the DIM successfully captures known historical dynamics of popular music, as validated using manually curated data. For example, it clearly shows the lineage leading from Reggae, Disco and Funk to modern electronic musical genres on one hand, and Hip hop and Rap on the other. The model also agrees with other measures of musical influence inferred from a large human curated musical website, `allmusic.com`. Finally, it reveals interesting connections between influence and innovation.

2. The Problem Setup

We first discuss the question of how to model musical influence. Then we present the dataset used to conduct this research.

2.1. How To Model Musical Influence

Influence relations in the corpus of popular music have complex structure, at several aspects. First, musical influence can be modeled at a hierarchy of levels, ranging from a sound segment – like an electronic distortion, to individual songs, to albums, to artists and musical bands. Second, the relation between these levels is “soft”: many songs are created in collaboration by several artists, many artists take part in several bands, and many songs were published in several versions, sometimes spanning a few decades. Finally, a well known thorny issue is that there exist no consistent metadata system which contains the above information for all music, and mapping music across metadata systems is hard.

With these considerations in mind, we chose to model influence on the basis of individual songs, since a song is typically a clearly delineated unit in terms of its acoustic data and metadata.

A second critical design choice is about the scope of influence. An artist may be influenced by another artist, or by an individual song. A single song may influence many artists, or even originate a musical style. Here we model influence as a process where one song affects the “musical language” of a musical stream, or “topic”. Such an approach was previously taken for modeling how one text document may influence an entire topic (Gerrish & Blei, 2010). This song-to-topic approach is expected to generalize better than direct song-to-song modeling, since it allows to control the model complexity by the number of topics.

This idea of song-to-topic influence hinges on the basic idea of topic modeling: each song has a distribution across a set of genres, and influences an entire topic (i.e. genre), in proportion to its membership in that topic. The goal of the model is to assign this song-level topic-influence score, and is described in detail in Section 3.

In our model we use only the acoustic data of a song, along with its year of release. We do not use any metadata such as genre, leaving this kind of information for validating out model.

Sevral previous works took different approaches to the question of measuing and modeling musical influence. Collins (2010) model influence as audio similarity, focusing on the influence of a single Depeche Mode song on several hundred British synth-pop songs. Later, the same authors (Collins, 2012) took a more elaborate approach to measure song influence, by building a probablistic model of the audio based on earlier songs, and evaluating the same model on later songs. Early songs that gave high likelihood for later songs were considered influential. This work focused on 248 from the early days of electronic dance music. Lastly, a different methodology was used by Bryan & Wang (2011). Using a large dataset indicating which songs sampled from which songs, they modeled influence as derived from the graph structure of which songs, artists, and genres sampled from whom. This method of course does not directly account for influence other than explicitely using an audio sample from an earlier song, a practice which is much more common in genres such as Hip hop and electronic music.

2.2. The Dataset

Conducting our study became possible with the publication of the *Million Songs Dataset* (Bertin-Mahieux et al., 2011) (MSD) in 2011. MSD is the first truly large scale, diverse and epoch-spanning dataset of songs ever made publicly available. MSD includes detailed audio features for $\sim$ 1,000,000 songs along with rich (albeit sometimes inconsistent and missing) metadata including genre tags, artist location and artist familiarity. The audio features are described in Section 4.

Out of the 1 million songs, about half are not tagged with a year of release, hence not suitable for our purpose. After removing these untagged songs along with

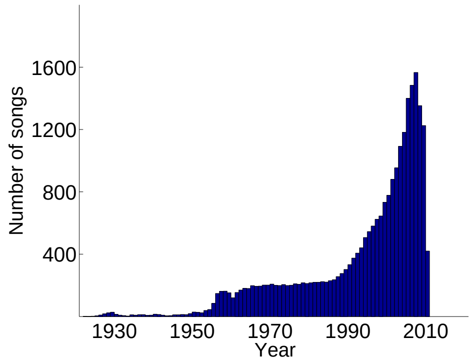

Figure 1. Number of songs per year in our dataset.

duplicates, very short songs, and items which are not musical in nature, such as comedy sketches, we are left with 455,123 songs from the years 1922-2010, with most songs from later years. In general only a few songs are available from any given album. We sampled 24941 songs, by 9222 artists. We biased our sample to include a relatively larger portion of earlier songs since our model revolves around modeling historical trends, and since the dataset itself is heavily skewed towards later year. We also biased our sample to include songs by more familiar artists, in order to aid non-musical-experts in appreciating some of the results. The ~25K songs were divided into 28 time epochs, with all songs of the same epoch treated as concurrent. We used 2-years epochs for the years 1963-2010, and longer epochs for earlier years, to compensate for the sparse data available before 1963. See Figure 1 for the distribution of songs over the years.

3. Modeling Influence

Contemporary music has a strong "topic-like" structure in the form of musical genres (like Hip hop, Metal, or Electronic), but at the same time, it exhibits nearly endless mixtures and interactions between genres. There is a clear sense of temporal evolution within and between these genres, which is fundamental to the modeling of influence (Holt, 2007; Fabbri, 1982).

To capture these structures and analyze the flow of musical influence across the music corpus, we use the tools of topic modeling. Specifically, we use the *Dynamic Topic Model* (DTM) (Blei & Lafferty, 2006) and the *Document Influence Model* (DIM) (Gerrish & Blei, 2010). Topic models provide a nuanced view of the structure of the musical corpus, enabling soft topic membership, and dynamic topic models have been shown to discover meaningful temporal structure in the evolution of heterogeneous texts (Blei & Lafferty, 2006; Hall et al., 2008).

Adopting these concepts to the evolution of music, we view influential songs as those songs which changed the "musical-language" of songs in their musical genres.

The model we use consists of three interacting layers, with inference performed jointly.

1. A classical topic model applied to each time epoch separately.
2. A *time-dependent* model: Each topic evolves with time, tying different epochs together.
3. A topic-dependent *influence* factor. Each song is seen as trying to "pull" future songs of its topic in its direction.

Formally, each song $d \in \{1 \dots D\}$ is comprised of a set of N_d *musical words*, $w_1^d, \dots, w_{N_d}^d$ taken from a vocabulary of size W. These words reflect both local and global audio structure, and are discussed in the next section. Each song belongs to one of T time epochs, and we assume the existence of K topics.

The topic model assigns a single topic k from $1 \dots K$ to the word w_n^d, and we indicate the assignment by an indicator variable $z_{n,k}^d$. The normalized sum $\frac{1}{N_d}\sum_{n=1}^{N_d} z_{n,k}^d$ is the proportion of topic k in song d.

In addition, we assign to each song a scalar normally distributed topic-influence score l_k^d controlling how much the topic k should later drift in direction of song d.

The following relations define the probabilistic model that we use:

For each epoch t and topic k the probability distribution of the words is governed by a W-dimensional parameter vector $\beta_{k,t}$.

The probability distribution is:

$$p(w|\beta_{k,t}(w)) \propto exp(\beta_{k,t}(w))$$

The temporal evolution of the topic-word distribution vectors β_t^k is given by:

$$\beta_{k,t+1}|\beta_{k,t} \sim \mathcal{N}\left(\mu_{k,t}, \sigma^2 I\right). \tag{1}$$

σ^2 controls the rate of the topics' evolution, and :

$$\begin{aligned}\mu_{k,t} &= \beta_{k,t}+ \\ &exp(-\beta_{k,t})\sum_d l_k^d \cdot \kappa(t,\tau(d))\sum_n w_n^d z_{n,k}^d.\end{aligned} \tag{2}$$

$z_{n,k}^d$ denote the topic-word assignments, l_k^d is each song's **topic-influence score**, $\tau(d)$ is the time of song

d and $\kappa(t, \tau(d))$ is a kernel function controlling the time-decay of the influence scores. Each epoch evolves from a starting point that is the sum of two components: the topic's distribution in the previous time-epoch, and the sum of the songs in the previous epochs, scaled by their influence score and a time-delay kernel.

Computing the posterior distribution of the topics and influence scores for this model is intractable. In their paper, Gerrish & Blei present a variational approximation and derive an algorithm for maximizing a lower bound on the marginal probability of the observed data. See Gerrish & Blei (2010) and the code at `code.google.com/p/princeton-statistical-learning`.

The variables of interest to us are the topic-influence scores l_k^d, and the aggregate topic-word assignments $\sum_n w_n^d z_{n,k}^d$. Together, they define the topic mixture of a song and how much it influences each of the topics. The influence of each song is defined as $l^d \equiv \max_k l_k^d$ (using the mean across topics gives similar results). We set the time-kernel κ to a log-normal distribution.

4. Features

Topic models were originally conceived for textual data, where each document is represented as a bag-of-words (Blei et al., 2003). Music however, is naturally represented as a single continuous variable, with structure on multiple time scales from less than a millisecond to the entire song length. To convert the continuous acoustic signals into a dictionary of discrete musical-signature, we applied a widely-used two-stage procedure: First, we extract short time-scale features on the scale of 0.25-1 seconds; then we quantize them using K-means. The clusters formed by K-means are treated as *musical-words*, and the histogram of their occurrence in a song gives us the required bag-of-words representation.

We compounded the short-scale audio features with long time-scale features such as tempo and key, and quantized these as well.

All of the raw audio features we used are available as part of the Million Songs Dataset, provided by The Echonest, and described in detail in the Echonest API documentation (Jehan, 2010).

More specifically, each song is partitioned into non-overlapping segments, typically under a second long, such that each segment is relatively uniform in timbre and harmony. For each segment we use 25 double precision features:

- **max. loudness**: A single number representing the peak dB value of the segment.
- **chroma**: A 12-component vector corresponding to the 12 pitch classes C, C#, D to B, with values ranging from 0 to 1 that describe the relative dominance of every pitch in the chromatic scale.
- **timbre**: A 12-component vector describing the quality of a musical note or sound that distinguishes different types of musical instruments, or voices. These are derived using the 12 top PCA components of a descriptor similar in nature to MFCC (Mermelstein, 1976).

These three types of features cover three primary and complementary musical facets (Serrà et al., 2012; Ball, 2010). Concatenating the chroma and timbre features together gives a richer description of each segment. It has been recently shown (Serrà et al., 2012) that while the overall distribution of chroma features is relatively stable over the years, the use of timbre features has been more variable. After concatenating and z-scoring the features, we applied K-means using $K = 5000$ to a set of 10 million descriptors, sampled randomly from songs in the MSD. Then, given a song, we make a hard-assignment of each segment of the song to one of the 5000 words, giving us a musical bag-of-words representation for each song with a vocabulary of 5000 musical-words.

In addition, we used 4 global audio variables:

- **tempo**: The overall estimated tempo of a track in beats per minute, usually associated with the speed or pace of the song. We quantized this variable into 12 words.
- **time signature**: The estimated meter of the song - how many beats are in each bar. Can take one of 7 discrete values.
- **key**: Identifies the tonic triad, the chord, major or minor, which represents the final point of rest of a piece. Can take one of 12 discrete values.
- **mode**: Indicates whether the song is major or minor. Can take one of 2 discrete values.

Overall, adding up the local and global features, we describe each song by a bag-of-words with a vocabulary size of **5033**. In the supplementary section we describe an experiment evaluating the contribution of each of the features.

5. Results

We applied 1, 5, 10 and 20-topic models to the 24941 songs data.

5.1. Comparison To Known Genres

We first looked at the matching of the topics with known musical genres. Our data includes 4803 genre tags, with a median number of 36 tags per artist. The genre-tags are weighted to indicate the strength of the genre-artist association. For each topic learned by our model, we summed up the artist-genre scores weighted by the topic proportions of each song. We found that the topics broadly match widely accepted genres such as Metal, Electronic, Hip hop, Indie Rock and Acoustic, especially for the later years where the dataset is larger and more varied.

We then investigated the temporal evolution of the topics, using the same genre scores. Table 1 shows the genre tags associated with 6 of the 20 topics of a 20-topic model, in select years from 1957 to 2009. We found several well known genre temporal dynamics. for example, a topic containing Jazz, Blues and Hard-bop songs in the late 50's evolved into a topic with Jazz, Funk, Disco and Soul in the 70's, Hip hop, Electro and Techno in the 80's, Hip hop, Electronica and Trip-hop in the 90's, and Techno and House in the 2000's. Note that these genre tags were in no way used in training or selecting the model.

5.2. Influential Songs

Table 2 shows some of the top ranking influential songs per topic and epoch. We chose examples that show diverse epochs and topics and highlight both well-known and lesser-known artists.

Table 3 presents several mistakes made by the model. We also consistently found that songs with wrong early time-stamps were given high influence scores by the model. For example, three 1992 songs by the Jamaican artist Barrington Levy were mislabeled as being from 1922. Every model we ran (until this mistake was discovered) found these songs to be extremely influential. This is because the model viewed them as predicting the acoustic language of their original, correct time, and scored them as influential.

5.3. Evaluating Against A Human Curated Influence Measure

The model we learned is unsupervised, and predicts the overall influence a song has had on all the songs which followed it.

To assess the validity and quantify the performance of our model, we compared our results with the database of the music site `allmusic.com`. This site includes encyclopedic data about ~100,000 artists, with a graph-like structure indicating artist-to-artist influenceas determined by the human editors of the site. For example, `allmusic.com` asserts that Beyoncé was influenced by Madonna, and that Dizzy Gillespie was influenced by Louis Armstrong.

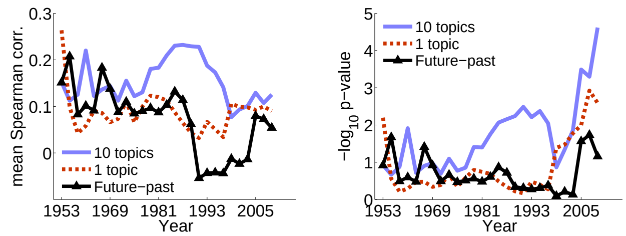

Figure 2. (left) The Spearman correlation and (right) the negative log p-value of the Spearman correlation across different epochs with `allmusic.com`'s influence rank for 10- and 1- topic models, and the future-past baseline described in 5.3. The highly significant p-values for the later years are possible because of the much larger number of songs available for those years - see Figure 1.

Our dataset consists of 24941 songs by 9222 artists; building the artist-to-artist influence graph for these artists using `allmusic.com`'s data, we found 5601 of these artists to have at least one edge in the graph.

To bring `allmusic.com`'s artist-to-artist relation and our song influence measures in line, we performed two procedures. First, we summed the number of artists each artist has influenced according to `allmusic.com`. Thus, according to this dataset we have the Beatles as the most influential artists with 556 influenced artists. Second, we had to transform our model from the song level to the artist level. We used a similar approach to that of (Bryan & Wang, 2011), averaging the influence scores of each artist, and yielding an artist influence score we denote l^{artist}_{mean}.

We found that according to the `allmusic.com` influence measure, earlier artists tend to be much more influential than later artists (having had time to acquire a larger following), making overall comparisons of influence mostly time related. We thus evaluated our influence measure seperately for each time-epoch, and averaged the results.

The mean Spearman rank correlation across epochs between the scores obtained from a 10-topic DIM and the `allmusic.com` data is 0.15 ($p < 0.05$). Figure 2 plots the per-epoch Spearman correlations and their respective negative log p-values for the 10- and 1-topic DIM, and a baseline method explained below. The 10-topic model performs best for songs from the mid 70's to early 90's, as well as the 2000's.

Figure 3 shows the mean Spearman correlations with `allmusic.com`'s data of 1, 5, 10 and 20-topic models, along with that of the best baseline model. We see that the 10-topic model performs best, and surprisingly better than the 20-topic model. This might stem from

	Topic #2	Topic #4	Topic #19	Topic #20	Topic #6	Topic #13
			2009-2010			
1	Electro	Hip hop	Indie	Indie rock	Folk	Heavy metal
2	Tech house	Rap	Folk	Indie	Ambient	Metal
3	Techno	Hardcore rap	Acoustic	Alternative	Chill out	Death metal
4	Electronica	Soul	Indie rock	Alternative rock	Jazz	Hardcore
5	Deep house	Reggae	Singer-songwriter	Punk	Indie	Metalcore
			1999-2000			
1	Hip hop	Hip hop	Jazz	Alternative rock	Jazz	Heavy metal
2	Downtempo	Rap	Folk	Indie rock	Classical	Metal
3	Electronica	Reggae	Hip hop	Punk	Downtempo	Death metal
4	Trip hop	Hardcore rap	Pop rock	Alternative	Folk	Punk
5	Electro	Gangster rap	Singer-songwriter	Indie	Folk rock	Thrash metal
			1989-1990			
1	Hip hop	Hip hop	Pop rock	Alternative rock	Jazz	Heavy metal
2	Electro	Pop rock	Classic rock	Hard rock	Smooth Jazz	Metal
3	Techno	Disco	Jazz	Pop rock	Easy listening	Thrash metal
4	Pop rap	Funk	Folk	Classic rock	Cool jazz	Death metal
5	Downtempo	Jazz	Blues	Heavy metal	Folk	Speed metal
			1979-1980			
1	Funk	Disco	Classic rock	New wave	Jazz	Punk
2	Jazz	Funk	Pop rock	Classic rock	Folk rock	New wave
3	Disco	New wave	Jazz	Pop rock	Singer songwriter	Metal
4	Reggae	Classic rock	Singer songwriter	80's	Soundtrack	Hard rock
5	Soul	Pop rock	Folk rock	Punk	Funk	Heavy metal
			1969-1970			
1	Classic rock	Classic rock	Classic rock	Classic rock	Classic rock	Classic rock
2	Jazz	Blues	Singer songwriter	Psychedelic rock	Jazz	Blues
3	Blues	Psychedelic rock	Folk rock	Blues	Singer songwriter	Hard rock
4	Pop rock	Jazz	Folk	Blues rock	Pop rock	Blues rock
5	Psychedelic rock	Blues rock	Blues	Pop rock	Blues	Psychedelic rock
			1957-1959			
1	Jazz	Jazz	Jazz	Classic rock	Jazz	Rockabilly
2	Hard bop	Blues	Blues	Jazz	Easy listening	Classic rock
3	Blues	Soul	Soul	Oldies	Vocal jazz	Doo-wop
4	Bebop	Oldies	Oldies	Rockabilly	Cool jazz	Oldies
5	Soul	Classic rock	Ballad	Blues	Smooth jazz	Rock 'n' roll

Table 1. Top genre tags for 6 topics from the 20-topic model, for sample time-epochs from 1957 to 2010. The topics were chosen to reflect several different genres. The genres come from artist metadata partly available in the Million Songs Dataset, and were not used in training or selecting the model.

the fact that the greater granularity of the 20-topic model prevents the model from identifying the globally most-influential artists, as they are represented in `allmusic.com`'s influence measure.

Future-Past Baseline

As a baseline for comparing the DIM performance, Gerrish & Blei (2010) proposed an easy to calculate heuristic influence measure. In this baseline, each word is assigned a weight for each time epoch by: $w_t = \frac{\text{Frequency of } w \text{ in } [t, t+f]}{\text{Frequency of } w \text{ in } [t-b, t]}$, where f and b denote the size of the time windows into the future and past, respectively. The influence of each song is taken as the mean over its word scores w_t as defined in above.We obtained a mean Spearman r of 0.07 with `allmusic.com`'s dataset ($p > 0.05$), maximized over all possible values of f and b. The correlations were only significant for the earliest and latest epochs, as shown in Figure 2.

6. Musical Innovation & Musical Influence

We use our model to explore the relation between being musically innovative and musically influential. This issue has only scarcely been looked at before (Noyes et al., 2010), and has never been approached using the audio content of the songs themselves.

Having established a valid computational model of musical influence, we are left to the task of modeling musical innovation. The DIM itself gives us a way to measure innovation. Since the model is probabilistic, each song is assigned a posterior likelihood. We propose to use this likelihood score as a measure of innovation,

Artist name	Song names	Year	Epoch influence rank in Topic #	Comments
Bob Dylan	Rainy Day Women #12 & 35, Like a Rolling Stone	1966, 1965	ranked 1^{st}, 2^{nd} in topic #16	Two of the most important songs by one of the most important musicians in popular music's history. These songs are part of Dylan's move to a more electric sound.
Killing Joke	The Wait	1979	ranked 1^{st} in topic #13	"Finding modest commercial success, Killing Joke have influenced many later artists, such as Nirvana, Nine Inch Nails, Metallica, Primus, Jane's Addiction, Soundgarden, Foo Fighters, Faith No More, and Rammstein, all of whom cited some debt of gratitude to Killing Joke." [Wikipedia]
Beastie Boys	Paul Revere	1986	ranked 1^{st} in topic #4	From the Beastie Boys debut album, "this album has been very influential since its debut" [urbandictionary.com]
Run-D.M.C.	Is It Live	1986	ranked 2^{nd} in topic #4, ranked 2^{nd} in topic #5	From their album Raising Hell: "The success of Raising Hell is often credited with kick-starting hip hop's golden age" [Wikipedia]
Elvis Presley	Mystery Train, Baby Let's Play House, That's All Right, I Got a Woman	1955, 1955 1954, 1956	ranked 1^{st}, 2^{nd}, 4^{th} & 5^{th} in topic #20	"Presley is regarded as one of the most important figures of 20th-century popular culture. He had a versatile voice and unusually wide success encompassing many genres" [Wikipedia]
Beck	Feather In Your Cap, Girl Dreams	1994	ranked 1^{st} in topic #8, ranked 1^{st} in topic #19	"...blurring boundaries and encapsulating how 90s hipsters looked toward the future" [allmusic.com]
Model 500 (Juan Atkins)	Night Drive Thru Babylon	1985	ranked 1^{st} in topic #2	"... is widely credited as the originator of techno music" [AOL music]. "At the dawn of the 1980s, Juan Atkins began recording what stands as perhaps the most influential body of work in the field of techno" [allmusic.com]

Table 2. A partial list of songs identified by our model as being the most influential songs for a specific topic and time, using the 20-topic model. We chose examples which illustrate a variety of genres and time-frames, and which present both well-known and lesser-known artists. Note that the same song may be influential in more than one topic.

Artist name	Song names	Year	Epoch influence rank in Topic #	Comments
Brook Benton	It's Just A Matter Of Time	1958	ranked 1^{st} in topic #20	This song ranked as most influential in this topic during this year, while the song and artist are considered not very significant. The artists that ranked just below him in this topic are considered much more influential: Johnny Cash, Hank Williams, Jimmy Reed and Chuck Berry.
Grand Funk Railroad	T.N.U.C.	1969	ranked 1^{st} in topic #9	Generally considered a band playing derivative music, "playing a loud, simple take on the blues-rock ... sound [allmusic.com]"

Table 3. A list of songs identified by our model as being among the most influential songs for a specific topic and time, using the 20-topic model. These examples illustrate errors of our model.

since more innovative songs will be harder to account for by the model, and thus assigned a lower likelihood. Innovation is of course always relative to the past, and so to measure the innovation of a song from 1960, we use a model fitted using only songs up to 1960. We will call this measure *time-restricted likelihood*.

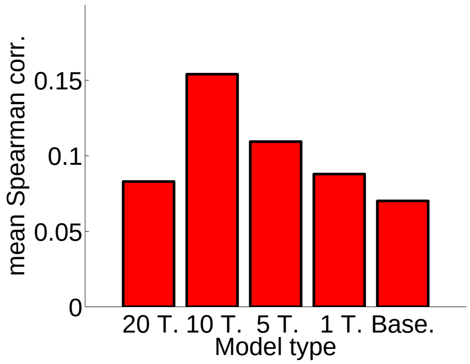

Figure 3. Mean per-epoch Spearman correlation with `allmusic.com`'s influence rank for 20-, 10-, 5- and 1- topic models, and the future-past baseline described in 5.3.

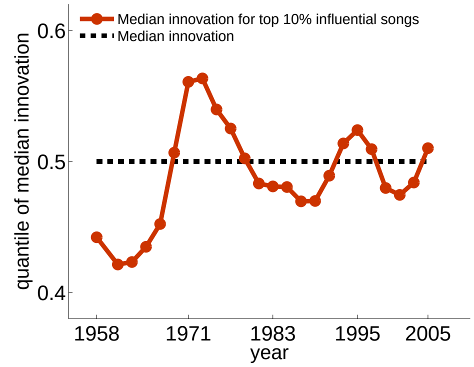

Figure 4. Median innovation of top 10% most influential songs, 1958-2005. In each epoch we standardize the median innovation to 0.5 (dashed black line). Median innovation for the most influential songs is below median at earlier years, and above median at the early 70's and mid 90's.

To validate that indeed *time-restricted likelihood* correlates with innovativeness, we conducted a qualitative survey on the single *least* likely songs from each time-epoch, as well as a comparable random selection of songs from the dataset. We found that 17 of 27 least likely songs are from artists or albums described as innovative or "experimental" during the relevant period: examples include songs from by Grandmaster Flash, considered by `allmusic.com` to be "Hip-hop's greatest innovator", and by the band Deerhoof, described as "an experimental spirit... challenging and utterly distinctive music"; unsuccessful examples include a song from Country singer Don Williams, "never known as an innovator". For a random song selection, we have found 8 out 27 can be considered innovative, 6 of them from the earlier periods of the dataset up to 1970. We also found time-restricted likelihood to correlate well with other measures of innovation such as the use of rarer musical-words relative to the epoch. We will thus refer to time-restricted likelihood as innovation score.

Before comparing influence to innovation, we have two dataset trends we need to consider. First, overall influence scores decline over the years. This results from the model being able to assign more credit for earlier songs as opposed to later songs, which have not yet had the chance to manifest their effect. Second, overall innovation scores increase with time. This is likely a result of the dataset including more songs and more diverse songs in later years. We account for these two trends by standardizing both the influence and innovation scores per each epoch, using order statistics. The most influential song in epoch t is given a normalized influence score of 1; the least influential a score of $\frac{1}{D_t}$, with D_t the number of songs in epoch t; the same goes for the innovation score.

After standardizing the influence and innovation scores, we find that overall across the dataset there is no monotonic correlation between the two (Spearman $r = -0.019, p > 0.05$). However, there are more subtle relations between the two. A closer look shows that the correlation fluctuates over the years. Figure 4 shows the median of the standardized innovation score for the top 10% most influential songs in each epoch. That is, we look at how innovative were the most influential songs, where innovation and influence are both measured relative to the period. We see two periods in which influential songs tended to be more innovative: the early 70's, and the mid-90's, and perhaps a third such period in the last few years. The rise at the mid-90's stems mainly from electronic and hip-hop artists who were given both high innovative *and* high influence scores; examples include Cypress Hill, Outkast, Tricky and Mad Professor. All are indeed considered both original and influential by critics.

7. Conclusion

We presented the first quantitative model of musical influence based on the sound content of popular songs. The learned influence scores are in agreement with manually curated data of artist-artist influence, providing a quantitative way, based on a principled probabilistic model, to study properties of the evolution of popular music.

Acknowledgments

We thank the anonymous reviewers for their helpful comments. We also thank Sean Gerrish and David Blei for making their code available, and Sean Gerrish for his assistance. U.S is a recipient of the Google Europe Fellowship in Machine Learning, and this research is supported in part by it. G.C. was supported by the Israeli science foundation Grant 1090/12, and by a Marie Curie reintegration grant PIRG06-GA-2009-256566

References

Ball, P. *The Music Instinct: How music works and why we can't do without it*. Random House, 2010.

Bertin-Mahieux, Thierry, Ellis, Daniel P.W., Whitman, Brian, and Lamere, Paul. The million song dataset. In *Proceedings of the 12th International Conference on Music Information Retrieval (ISMIR 2011)*, 2011.

Blei, D.M. and Lafferty, J.D. Dynamic topic models. In *Proceedings of the 23rd international conference on Machine learning*, pp. 113–120. ACM, 2006.

Blei, D.M., Ng, A.Y., and Jordan, M.I. Latent dirichlet allocation. *the Journal of machine Learning research*, 3:993–1022, 2003.

Bloom, H. *The anxiety of influence: A theory of poetry*. Oxford University Press, USA, 1997.

Bryan, Nicholas J and Wang, Ge. Musical influence network analysis and rank of sampled-based music. In *Proceedings of the 12th International Society for Music Information Retrieval Conference*, 2011.

Collins, Nick. Computational analysis of musical influence: A musicological case study using mir tools. In *Proceedings of the International Symposium on Music Information Retrieval*, 2010.

Collins, Nick. Influence in early electronic dance music: an audio content analysis investigation. In *Proceedings of the International Symposium on Music Information Retrieval*, 2012.

Fabbri, F. A theory of musical genres: Two applications. *Popular Music Perspectives*, 1:52–81, 1982.

Gerrish, S. and Blei, D.M. A language-based approach to measuring scholarly impact. In *Proceedings of the 26th International Conference on Machine Learning*, 2010.

Hall, David, Jurafsky, Daniel, and Manning, Christopher D. Studying the history of ideas using topic models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 363–371, 2008.

Holt, F. *Genre in popular music*. University of Chicago Press, 2007.

Jehan, T. The Echo Nest Analyze documentation. Technical report, The Echo Nest, 2010.

Lethem, J. The Ecstasy of Influence. *Harper's Magazine*, pp. 59–71, 2007.

Logan, B. Content-based playlist generation: Exploratory experiments. In *Proceedings of the International Symposium on Music Information Retrieval*, 2002.

Logan, B. Music recommendation from song sets. In *Proc ISMIR*, pp. 425–428. Citeseer, 2004.

Lu, L., You, H., Zhang, H.J., et al. A new approach to query by humming in music retrieval. In *Proceedings of the IEEE International Conference on Multimedia and Expo*, 2001.

McFee, B. and Lanckriet, G. R. G. The natural language of playlists. In *12th International Symposium for Music Information Retrieval (ISMIR2011))*, October 2011.

McFee, B., Barrington, L., and Lanckriet, G.R.G. Learning content similarity for music recommendation. *IEEE Transactions on Audio, Speech, and Language Processing*, 20(8):2207–2218, October 2012.

Mermelstein, P. Distance measures for speech recognition, psychological and instrumental. *Pattern recognition and artificial intelligence*, 116:91–103, 1976.

Noyes, E., Allen, I.E., and Parise, S. Artistic influences and innovation in the popular music industry. *Frontiers of Entrepreneurship Research*, 30(15):3, 2010.

Reynolds, S. *Retromania: Pop Culture's Addiction to Its Own Past*. Isbn Edizioni, 2011.

Scaringella, N., Zoia, G., and Mlynek, D. Automatic genre classification of music content: a survey. *Signal Processing Magazine, IEEE*, 23(2):133–141, 2006.

Serrà, J., Corral, Á., Boguñá, M., Haro, M., and Arcos, J.L. Measuring the evolution of contemporary western popular music. *Scientific Reports*, 2, 2012.

Su, J.H., Yeh, H.H., Yu, P.S., and Tseng, V.S. Music recommendation using content and context information mining. *Intelligent Systems, IEEE*, 25(1):16–26, 2010.

Turnbull, D.R., Barrington, L., Lanckriet, G., and Yazdani, M. Combining audio content and social context for semantic music discovery. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, pp. 387–394. ACM, 2009.