Temporal and Regional Variation in Rap Lyrics

Christopher Johnson-Roberson
Brown University
Providence, RI 02912
christopher_johnson-roberson@brown.edu

Matthew Johnson-Roberson
University of Michigan
Ann Arbor, MI 48109
mattjr@umich.edu

Abstract

Probabilistic topic models offer a way to explore rap lyrics across space and time, on a larger scale than is possible in traditional content analysis. By applying latent Dirichlet allocation (LDA) and Dirichlet-multinomial regression (DMR) to a corpus of transcribed rap lyrics, we uncover expected topics such as street life, drugs, and violence, but also less obvious ones such as "family/childhood reminiscence" which have not been adequately addressed in the literature. We incorporate time and location metadata into the model, both to improve model quality and to display the temporal and regional distributions of these themes. This work challenges the narrative of a unilateral shift from an abstract "ghetto" to a concrete "hood" in rappers' conceptions of place.

1 Introduction

Since its genesis in the late 1970s, rap music's sonic and lyrical content have been closely linked with its geospatial and historical context. Scholars have connected rap with the landscape of the postindustrial city [1] and speculated about discursive shifts in rappers' relationship to place, such as a purported move from the generalized "ghetto" to the specific, concrete "hood" [2]. Despite broad interest in content analysis of rap's overarching themes, extant studies have typically focused on single aspects of the genre and incorporated 500 or fewer songs due to the constraints of manual analysis.

Probabilistic topic models [3,4] enable exploration of text corpora on an unprecedented scale. Earlier work has used topic models to examine the history of ideas in scientific fields [5] and regional variation in language use on Twitter [6], but no such studies have yet addressed rap music and the spatiotemporal variation of its lyrical themes. Topic models can both chart the trajectories of known themes within rap, and highlight other themes and trends that might otherwise be overlooked. We utilized latent Dirichlet allocation (LDA) [3] to generate an initial topic model of rap lyrics; posterior predictive checking [7] allowed us to evaluate the model and determine which metadata features had the greatest effect on the formation of its topics. We then employed Dirichlet-multinomial regression (DMR) [8] to generate topics conditioned on the times, places, and artists in the corpus.

2 Data and Methods

Lyrics were downloaded from the Original Hip-Hop Lyrics Archive (`http://www.ohhla.com`), a user-submitted lyrics site. Of the 34492 text files obtained, 18149 (52%) had both viable location data from the Echo Nest API (`http://developer.echonest.com/`) and album release dates from the Spotify Metadata API (`https://developer.spotify.com/technologies/web-api/`), spanning the years 1983-2013.[1]

[1] All the Python and R scripts used to generate this analysis are available at `http://www.github.com/chrisjr/ohhla_analysis`.

We trained LDA and DMR models using MALLET [9], an open-source machine learning package. Texts were preprocessed to eliminate terms that occurred in fewer than 3 documents; a tf-idf (term frequency-inverse document frequency) filtering scheme reduced the number of word types to no more than 5000. The resulting corpus has 4826 word types, for a total of 2.1 million tokens; it includes 1322 artists, 500 places, and 3825 typists (transcribers of lyrics).

LDA models were trained with the number of topics K ranging from 25 to 150. After inspecting the results, we chose to use 50 topics which provided the best trade-off between interpretability and specificity. Each model ran for 1000 iterations, with a burn-in period of 200 iterations and hyperparameter optimization every 10 iterations afterward.

We plotted the words and ranks of several topics learned by the model, focusing on those in the interquartile range of mutual information of words W and documents D given topic k (denoted as $MI(W, D|k)$). One immediate application was to test the assertion that rap lyrics shifted from discussing a generalized "ghetto" to a more specific "hood"; thus, we also graphed the prevalence of those terms in the $\{hood, ghetto, streets\}$ topic as a proportion of the total number of tokens in each year, and their instantaneous mutual information (IMI) with groupings by artist, place, and time.

After constructing the initial (metadata-agnostic) model, we evaluated it for systematic variation over different types of metadata. Following the posterior predictive check developed by Mimno and Blei [7], we generated 100 replications of the model's posterior distribution based on its Gibbs state, replacing each word with a new one sampled from the multinomial distribution of its topic. We then compared different subsets of the corpus based on place, year, artist, and typist for mutual information between groupings and topics; rather than plotting the MI directly, we calculated its deviance from the replicas $\frac{observed\ MI - mean\ replicated\ MI}{std.dev.\ of\ replicated\ MIs}$. This gave a clear indication of which metadata features were most influential on the topics learned by the model.

Once we found that artist, place, and year had the greatest effect on the induced topics, we fitted a Dirichlet-multinomial regression model [8] to the data (also using MALLET), with indicator variables for place and artist and the sufficient statistics $\log(p_d)$ and $\log(1 - p_d)$ of continuous variable $p_d = \frac{year_{doc} - year_{min}}{year_{max} - year_{min}}$ for time.

Finally, we overlaid 2D kernel density estimation plots of the regional prevalence of several topics onto maps of the United States, in order to visualize geographic trends in topic distribution.

3 Results

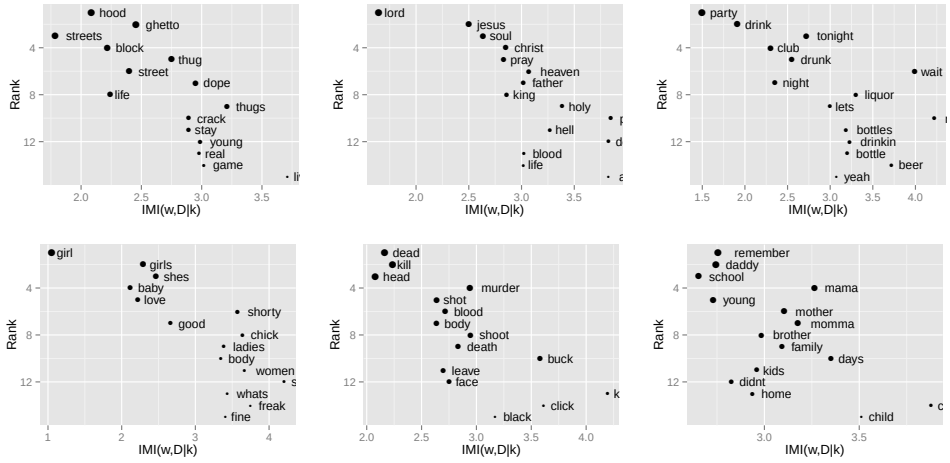


Figure 1: **Sample topics from the LDA model.** The y-axis shows term rank in the topic, and the x-axis shows the instantaneous mutual information of the word and set of documents given the topic (a measure of "specificity" of the word to the documents, as described in [7]).

Many topics evinced clear semantic content and were familiar from experience with the genre; for example, those shown in Figure 1 appear to refer to the ghetto/hood, religion, drinking/partying, women, violence, and family/childhood reminiscences.

Examining the relative prevalence of "ghetto" and "hood" in the $\{hood, ghetto, streets\}$ topic, (Figure 2), we found that both terms appeared fairly frequently in the topic throughout. Despite earlier claims of a marked discursive shift between the two terms [2], their co-presence in this topic suggests that both words have continued to be deployed in similar contexts over time. This provides empirical support for the intuitive commonality between these terms described in [10].

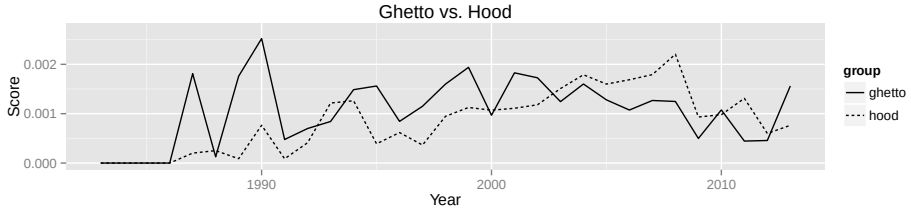


Figure 2: **Shift in usage from "ghetto" to "hood" is later and less dramatic than previously asserted.** Contrary to claims that references to the "hood" increased rapidly starting in 1987-1988 [2] or fell off after the 1990s [10], we find that "hood" has been on the rise since the early '90s and came into the lead around 2003.

Grouping the documents according to artist, place, and year, we calculated the instantaneous mutual information of the top terms in this topic with respect to each grouping. Figure 3 shows that "ghetto" is actually more particular than "hood" in its usage by specific artists and in specific places, but more general in relation to time.

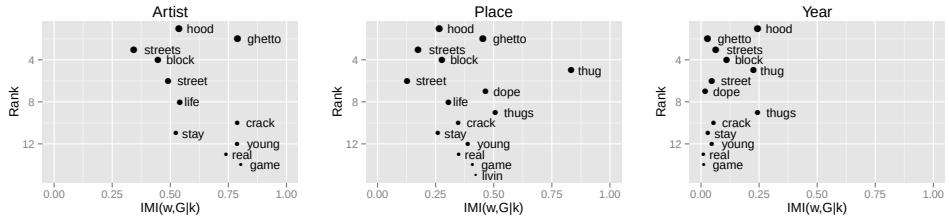


Figure 3: **Use of "ghetto" vs. "hood" by metadata category.** The use of "ghetto" is more particular to certain artists and places than "hood," but is more general in terms of time.

We compared the observed discrepancy with the mean replicated discrepancy for each topic over a variety of groupings, finding that artist, place, and year were most strongly associated with systematic deviations from the overall expected topic distributions (Figure 4). The five most deviant topics when grouping by artist are shown in Table 1.

In accordance with the findings shown in Figure 4, we fed place, time, and artist features into Dirichlet-multinomial regression [8] to improve model quality. We used the Gibbs state from the DMR model to produce the topic-place counts for our 2D kernel density estimates, illustrating the geographical distributions of the topics (Figure 5).

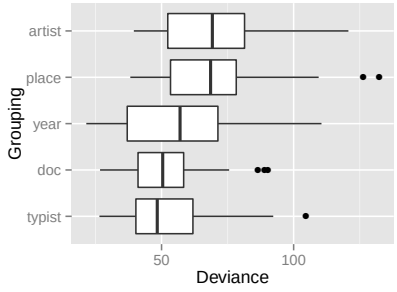
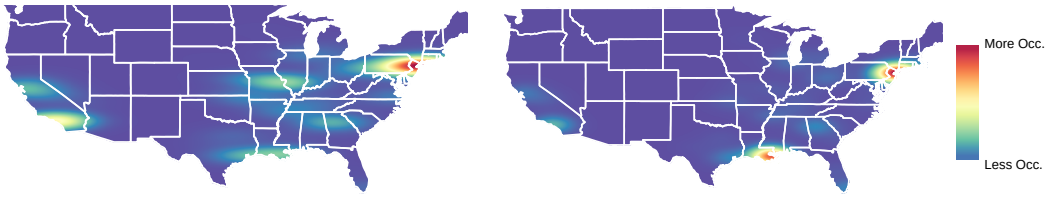


Figure 4: **Systematic variation over metadata features.** Artists gave greater than expected information about topic assignments, followed by places and years; typists generally gave no more information about topics than did documents.

Table 1: **Artist-topic pointwise mutual information.** Deviance is the deviation compared to the Gibbs state replicas when grouping by artists, as in Figure 4. Normalized pointwise mutual information ($npmi$) indicates the strength of artist-topic associations; most are predictable (including artists such as Tech N9ne referring to themselves, or the association between Spanish words and Spanish-language rappers), while others (such as $\{rock, rockin, heavy\}$, connected with a number of Jamaican artists and one Finnish group) suggest areas for further investigation.

Topic	**Deviance**	**Top Artists** $(npmi)$[2]
$\{bang, tech, n\}$	120.73	Tech N9ne (0.48), Body Count (0.37), Sway & Tech (0.37)
$\{funky, movin, rollin\}$	113.79	SWV (0.45), DJ Tomekk (0.4), George Clinton (0.37)
$\{como, vida, para\}$	105.56	Dyablo (0.8), Daddy Yankee (0.73), Control Machete (0.73)
$\{check, word, crew\}$	96.56	Grand Puba (0.36), Das EFX (0.34), Nefertiti (0.3)
$\{rock, rockin, heavy\}$	93.61	Super Cat (0.5), Bomfunk MC's (0.41), Sean Paul (0.4)



(a) $\{gangsta, gang, gangstas\}$ (b) $\{ghetto, thug, hood\}$

Figure 5: **Geographical specificity of themes.** The topic $\{gangsta, gang, gangstas\}$ is widespread across the country, albeit with increased intensity in the Midwest. $\{ghetto, thug, hood\}$ reveals a localized peak in New Orleans, prompted by artists such as Master P, Silkk the Shocker, and Mystikal who frequently reference "soldiers."

4 Discussion

The topic models successfully learned a number of discourses in the genre that are easily recognizable to scholars and fans: many of those seen in Figure 1, as well as others such as marijuana smoking and crass and misogynist language. Other topics are less obvious, yet still recognizable and coherent: the topic about family and childhood is easily recognizable to hip-hop fans but has rarely been discussed by scholars in a positive light. Although this topic comprises a large proportion of well-known songs such as 2Pac's "Dear Mama," it may have been left out of prior content analyses due to their varied and generally negative emphases, such as nihilism [11], misogyny [12], and violence [13] in the genre. The wide net cast by topic modeling can alert us to themes that would otherwise escape notice, directing our attention to understudied areas and helping to clarify what we know about recognizable tropes.

5 Conclusions and Future Directions

The use of probabilistic topic models enabled exploration of large text corpora, such as the Original Hip-Hop Lyrics Archive's collection of transcribed lyrics. LDA and DMR facilitated discovery of underexplored themes, such as the "family/childhood" topic, and served to show the diffusion of known themes over space and time. Topic modeling also revealed the similar usage contexts and timeframes of "ghetto" and "hood," contrary to the notion that these terms evince radically different ideas of place. In fact, mutual information between words and document groups suggested that "hood" may be the more broadly used term across artists and locales.

[2]The pointwise mutual information $pmi(a, k)$ is defined as $\log \frac{p(a,k)}{p(a)p(k)}$, where $p(a, k)$ is the number of tokens assigned to a given artist and topic divided by the total number of tokens. The parenthetical values by each artist in the table are equal to $npmi(a, k) = \frac{pmi(a,k)}{-\log p(a,k)}$, which ranges from 0 (no mutual information) to 1 (always coinciding).

Much of rap's meaning derives from the way in which the lyrics are delivered, as well as the interplay of music and text; indeed, it has been argued that since the "beat" (musical backing) often precedes the writing and delivery of the lyrics, sonic and musical aspects should assume primary importance in analysis [14]. To this end, we might utilize the acoustic features analyzed by The Echo Nest (e.g. tempo, timbral characteristics, etc.), or devise a generative model that jointly models acoustic [15] and linguistic features, in order to understand better how sound relates to rap's lyrical content.

References

[1] Tricia Rose. *Black noise: rap music and black culture in contemporary America*. University Press of New England, Hanover, NH, 1994.

[2] Murray Forman. "Represent": race, space and place in rap music. *Popular Music*, 19(01):65–90, 2000.

[3] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

[4] David M. Blei. Probabilistic topic models. *Communications of the ACM*, 55(4):77–84, 2012.

[5] David Hall, Daniel Jurafsky, and Christopher D. Manning. Studying the history of ideas using topic models. In *EMNLP '08*, pages 363–371, 2008.

[6] Jacob Eisenstein, Brendan O'Connor, Noah A. Smith, and Eric P. Xing. A latent variable model for geographic lexical variation. In *EMNLP '10*, pages 1277–1287, 2010.

[7] David Mimno and David Blei. Bayesian checking for topic models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 227–237, 2011.

[8] David Mimno and Andrew McCallum. Topic models conditioned on arbitrary features with dirichlet-multinomial regression. In *UAI*, 2008.

[9] Andrew McCallum. Mallet: A machine learning for language toolkit. http://mallet.cs.umass.edu, 2002.

[10] Rivke Jaffe. Talkin' 'bout the ghetto: Popular culture and urban imaginaries of immobility. *International Journal of Urban and Regional Research*, 36(4):674–688, 2012.

[11] Charis E. Kubrin. "I See Death Around the Corner": Nihilism in rap music. *Sociological Perspectives*, 48(4):433–459, December 2005.

[12] Ronald Weitzer and Charis E. Kubrin. Misogyny in rap music: A content analysis of prevalence and meanings. *Men and Masculinities*, 12(1):3–29, October 2009.

[13] Denise Herd. Changing images of violence in rap music lyrics: 1979–1997. *Journal of Public Health Policy*, 30(4):395–406, 2009.

[14] Kyle Adams. Aspects of the music/text relationship in rap. *Music Theory Online*, 14(2), May 2008.

[15] S. Kim, S. Narayanan, and S. Sundaram. Acoustic topic model for audio information retrieval. In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, 2009. WASPAA '09*, pages 37–40, 2009.