

From Structure to Substance: A Network and Content Analysis of the Hip-Hop Industry

Mohammed Pathariya (mjpathar@iu.edu) November 7, 2025

1. Introduction and Motivation

The hip-hop industry, a dominant force in global culture, is often discussed in terms of its distinct "scenes" or "communities" (e.g., West Coast, Atlanta, Chicago). However, these labels are often impressionistic. Are these communities real, data-driven clusters? And if so, are they defined merely by social connections, or do they share unique thematic identities? This project seeks to answer these questions by moving beyond casual observation to computationally quantify and map the hip-hop industry.

My core research question is twofold:

1. What is the large-scale social *structure* of the hip-hop industry, based on artist collaborations?
2. Do these structural communities also share distinct *thematic* content, as evidenced by their lyrics?

This project is significant because it aims to create a multi-layered "blueprint" of a massive cultural domain. It will synthesize two distinct computational methods—network science and natural language processing—to explore the alignment between social structure and thematic content. The findings will provide a data-backed framework for understanding how artistic scenes form, cohere, and differentiate themselves.

2. Light Literature Review

This research is grounded in two primary fields. The first is **network science**, which provides the tools to model and analyze large-scale social structures. For this, I will rely on established methods for community detection, such as the Louvain method (Blondel et al., 2008), which is a highly efficient algorithm for discovering dense clusters in large networks. This method will be used to identify the "communities" of artists based on their collaboration patterns.

The second field is **computational content analysis**, specifically **probabilistic topic modeling**. The foundational algorithm for this is Latent Dirichlet Allocation (LDA) (Blei et al., 2003). LDA is an unsupervised machine learning technique that automatically discovers the abstract "topics" (clusters of co-occurring words) present in a large collection of documents. This method will be used to analyze the lyrical corpus.

This project's novelty lies in its synthesis of these two methods. It will not just identify structural communities (per Blondel et al.) or thematic topics (per Blei et al.), but will analyze the *alignment* between them, providing a richer, "3D" view of a cultural field.

3. Proposed Methodology

This project is a two-part capstone that will re-use and build upon the data collection methods defined in Assignment 2, while applying a "distinct analysis" for Assignment 3, as required.

- **Data Source:** The project will use a novel data source: the **Genius.com API**. A successful feasibility study has already been completed, confirming that I can access all required data points.
- **Phase 1: Data Collection (The Network)** I will write a Python script using the lyricsgenius library to perform a "1.5-degree" data crawl. I will begin with a seed set of 3-5 influential artists (e.g., Kendrick Lamar, Drake, J. Cole) representing different scenes. The script will find all of their songs and then scrape the **featured_artists** list from each song. The final network dataset will consist of this seed set plus all of their direct collaborators.
- **Phase 2: Analysis (Network Science)** The data from Phase 1 will be used to construct a collaboration network, where **Nodes = Artists** and **Edges = A collaboration on a song**. I will then perform a network analysis to:
 1. Identify central artists ("super-connectors") using PageRank and Betweenness Centrality.
 2. Discover the structural "scenes" using Louvain community detection. This analysis will form the basis of the Assignment 2 report.
- **Phase 3: Data Collection (The Content)** Concurrently with Phase 1, the data collection script will also scrape and save the full **lyrics** for every song in the network. This text corpus is the new, distinct dataset that will be used for the capstone analysis.
- **Phase 4: Analysis (NLP & Synthesis)** This is the "distinct analysis" for Assignment 3.
 1. **Preprocessing:** I will clean the entire lyrical corpus using a standard NLP pipeline (e.g., removing stopwords, lemmatization).
 2. **Topic Modeling:** I will train an LDA topic model on the cleaned corpus to discover 20-30 distinct lyrical "topics" (e.g., a topic for "materialism" identified by words like money, cars, ice, and a topic for "social consciousness" with words like struggle, police, system).
 3. **Synthesis:** This is the core of the capstone. I will merge the results from A2 and A3. I will analyze the topic distribution for each network community. This will answer my main research question: "Does the 'TDE' network community (from A2) have a statistically significant-higher focus on the 'social consciousness' topic (from A3) than the 'Atlanta' community?"

4. Feasibility and Expected Findings

This project is highly feasible. The successful feasibility study confirmed that the lyricsgenius library can reliably retrieve both **featured_artists** (for A2) and **lyrics** (for A3), which are the two critical data points. The scope is manageable by limiting the crawl to a 1.5-degree network from a few seed artists.

I expect to find a strong correlation between network structure and thematic content. My hypothesis is that the algorithmically-discovered network communities will align cleanly with distinct lyrical topics. For example, I predict the "Atlanta" community will show a high prevalence of "materialism" topics, while the "TDE/Dreamville" communities will show a higher prevalence of "social consciousness" or "storytelling" topics. This project will provide quantitative, data-driven evidence for the existence and unique identities of different "scenes" in hip-hop.

5. References

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3(Jan), 993-1022.

Blondel, V. D., Guillaume, J. L., Lambiotte, R., & Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10), P10008.