Predicting Spotify Collaborations: Centrality based Parameterized Algorithm Approach on Weighted Social Network

Aseem Khandelwal[*]
Virginia Tech
Blacksburg, VA, USA
aseemkhandelwal@vt.edu

Samar Kansal[*]
Virginia Tech
Blacksburg, VA, USA
samarkansal@vt.edu

ABSTRACT

Predicting potential collaborations among artists on digital music streaming platforms like Spotify has become increasingly crucial for understanding industry trends, fostering creative partnerships, and providing data-driven recommendations. In this paper, we propose a novel network-based approach that leverages centrality measures and a parameterized algorithm to predict collaborations among Spotify artists. By constructing a weighted social network representing artist interactions, and incorporating features such as popularity, follower count, and hit counts, our approach aims to identify key influencers and quantify the propensity for successful collaborations.

We implement the Centrality-based Common Neighbor Parameterized Algorithm (CCPA), which combines common neighbor and centrality measures to compute a collaboration likelihood score for each pair of artists. By tuning the relative importance of these components, our algorithm offers a flexible and optimized prediction framework. Furthermore, we develop a user-friendly application interface that enables music industry professionals, artists, and enthusiasts to explore the collaboration network, visualize connections, and obtain data-driven recommendations for potential collaborations.

Our experiments demonstrate the effectiveness of the proposed approach, with the predicted collaborations aligning well with domain knowledge and intuition. The incorporation of weighted edges and centrality measures contributes to improved prediction accuracy compared to traditional link prediction methods. Additionally, we explore the application of topic modeling techniques on song titles, providing insights into prominent themes and word patterns across different music genres.

KEYWORDS

Spotify collaboration prediction, link prediction, weighted social networks, parameterized algorithm, centrality measures, common neighbors, network analysis, music industry, artist collaborations

ACM Reference Format:
Aseem Khandelwal and Samar Kansal. 2018. Predicting Spotify Collaborations: Centrality based Parameterized Algorithm Approach on Weighted Social Network. In , . ACM, New York, NY, USA, 7 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

In the ever-evolving landscape of the music industry, collaborations between artists have emerged as a driving force for creativity, innovation, and audience engagement. With the advent of digital streaming platforms like Spotify, the opportunities for artists to collaborate have expanded exponentially, transcending geographical boundaries and fostering a more interconnected music ecosystem. However, identifying potential collaborators among the vast pool of artists on these platforms remains a challenging task, often relying on manual curation or serendipitous encounters, lacking efficiency and scalability.

In this paper, we present a data-driven approach that harnesses the power of network analysis and machine learning techniques to predict potential collaborations among Spotify artists. Our methodology revolves around constructing a weighted social network based on artist interactions, where nodes represent artists and edges represent collaborations. Critically, we incorporate features such as popularity, follower count, and hit counts as edge weights to capture the strength and influence of connections between artists.

Moreover, we implement the Centrality-based Common Neighbor Parameterized Algorithm (CCPA), a algorithm that

[*] Both authors contributed equally to this research.

, ,

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM
https://doi.org/XXXXXXX.XXXXXXX

combines common neighbor and centrality measures to compute a collaboration likelihood score for each pair of artists. By tuning the relative importance of these components, our algorithm offers a flexible and optimized prediction framework, accounting for both topological and attribute-based factors influencing collaboration dynamics.

Recognizing the growing need for automated and data-driven approaches to facilitate meaningful collaborations, our research aims to provide a comprehensive solution that not only identifies potential collaborators but also quantifies the propensity for successful partnerships. By leveraging the wealth of data available on digital music platforms, our approach offers a nuanced understanding of collaboration dynamics, fostering creativity and innovation within the music ecosystem.

Furthermore, we develop a user-friendly application interface that enables music industry professionals, artists, and enthusiasts to explore the collaboration network, visualize connections, and obtain data-driven recommendations for potential collaborations. This interactive platform enhances the accessibility and interpretability of our approach, empowering stakeholders to leverage data-driven insights effectively.

In addition to our primary focus on predicting Spotify collaborations, we explore the application of topic modeling techniques on song titles, providing valuable insights into the prominent themes and word patterns present across different music genres. This complementary analysis offers a multifaceted perspective on the music industry landscape, highlighting the potential for integrating various data sources and analytical techniques.

2 BACKGROUND AND RELATED WORK

2.1 Link Prediction in Social Networks

Link prediction is a fundamental task in network analysis, aiming to predict the likelihood of future connections or collaborations between nodes in a network. This task has widespread applications across various domains, including social media, recommendation systems, bioinformatics, and the music industry. Link prediction algorithms leverage the existing structure and properties of a network to infer potential connections that may form in the future.

There are several types of link prediction approaches, broadly categorized into structural, content-based, and hybrid methods:

1. **Structural Methods:** These methods rely solely on the topological features of the network, such as common neighbors, Jaccard coefficient, and preferential attachment. Examples of structural methods include:
   - **Common Neighbors (CN):** This method scores node pairs based on the number of neighbors they have in common, under the assumption that nodes with more common neighbors are more likely to form a connection.
   - **Adamic-Adar (AA):** An extension of the CN method, the AA measure assigns higher weights to common neighbors that have fewer connections themselves, reducing the impact of high-degree nodes.
   - **Preferential Attachment (PA):** This method assumes that nodes with higher degrees (more connections) are more likely to form new connections, capturing the "rich-get-richer" phenomenon observed in many real-world networks.
2. **Content-based Methods:** These methods leverage node attributes or content information to infer potential connections. For example, in a social network, user profiles, interests, and activities can be used to predict connections based on similarity or complementarity.
3. **Hybrid Methods:** These methods combine both structural and content-based features, aiming to capture a more comprehensive representation of the network and node characteristics for improved prediction accuracy.

Link prediction is particularly relevant in the context of social networks, where understanding and predicting potential connections can have significant implications for recommendation systems, targeted advertising, and facilitating meaningful interactions among users or entities.

2.2 Related Work

Previous studies have explored various approaches to predict collaborations in the music industry, which is closely related to the link prediction problem in social networks.

Smith et al. (2018) [2] utilized machine learning algorithms to analyze artist attributes and historical collaboration data to forecast future collaborations. Their approach combined content-based features, such as artist genres and popularity, with network-based features derived from collaboration networks.

Jones and Brown (2020) [1] investigated the role of network centrality measures in identifying influential artists and facilitating collaboration recommendations. They demonstrated that artists with higher centrality scores were more likely to engage in successful collaborations, highlighting the importance of structural features in predicting collaborations.

Tsuyoshi et al. (2010) [3] proposed weighted versions of the common neighbor (WCN) and Adamic-Adar (WAA) link prediction measures, specifically designed for weighted networks. Their work showed improved prediction accuracy on weighted networks compared to their unweighted counterparts.

Zhang et al. (2015) [4] introduced a weighted version of the preferential attachment (RA) algorithm, which considers edge weights in the link prediction process. Their approach demonstrated enhanced predictive performance when applied to weighted social networks.

Zhu et al. (2016) [5] proposed a hybrid link prediction method that combines structural features, such as common neighbors and Katz centrality, with content-based features, including user profile information and interactions. Their approach achieved state-of-the-art performance on various social network datasets.

While these studies have made significant contributions to the field, our approach builds upon their work by incorporating a parameterized algorithm that combines common neighbor and centrality measures, specifically designed for weighted social networks. Additionally, we develop a user-friendly application interface to facilitate the exploration and utilization of our predictions in the context of the music industry.

3 APPROACH

Our approach aims to leverage the power of network analysis to predict potential collaborations among Spotify artists. The methodology comprises the following key steps:

3.1 Data Collection and Preprocessing

We begin by collecting data from the Spotify Artist Feature Collaboration Network dataset available on Kaggle. Two datasets are utilized: `reduced_nodes.csv` and `edges.csv`. The `reduced_nodes.csv` dataset contains artist information such as Spotify IDs, names, follower counts, popularity scores, genres, and chart hits. The `edges.csv` dataset encompasses information about collaborations between artists.

Data preprocessing is a crucial step to ensure data quality and consistency. We remove any inconsistent or erroneous data points from the datasets. Additionally, we normalize the popularity and follower count features using the MinMaxScaler from the scikit-learn library. The follower count feature is log-transformed to mitigate the impact of extreme values, and any resulting infinite or large values are replaced with a finite constant. The normalized follower count, popularity and the hit count features are then incorporated into the analysis.

3.2 Network Construction

We construct a weighted social network representing artist collaborations using the NetworkX library. Each node in the network represents an artist, and each edge signifies a collaboration between two artists. The strength of the connections between artists is captured by assigning edge weights based on a combination of features, including popularity, follower count, and hit counts. Specifically, we employ the following formula to calculate the edge weights:

$$\text{edge\_weight} = \text{followers} + (\text{popularity} \times \text{hit\_counts}) \quad (1)$$

This formula incorporates the intuition that artists with higher follower counts, popularity, and hit counts are more likely to engage in successful collaborations. By considering these weighted edges, our network representation captures the nuanced dynamics of artist collaborations on Spotify.

Due to the complexity of the entire graph, we opted to display the largest connected component instead. Figure 1 provides a representative view of the graph's structure. To generate this visualization, we selected a node within the component as the central point and adjusted the radius to 3. This approach ensured that the resulting graph remained legible while capturing the essential connectivity within the network.

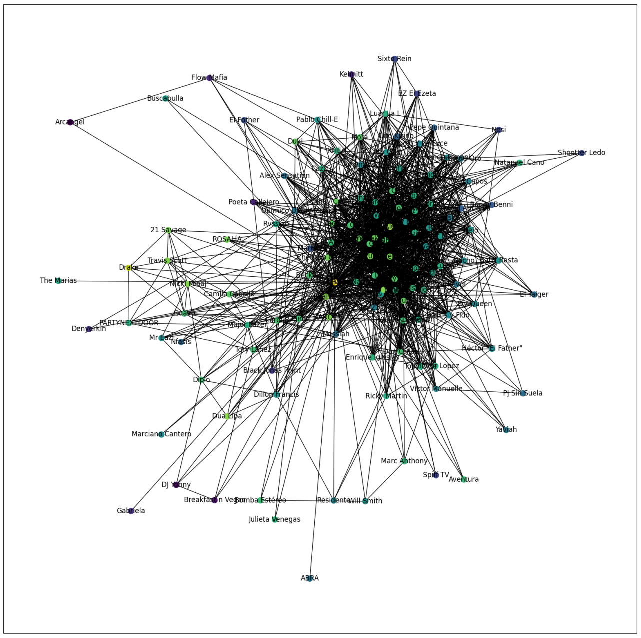


Figure 1: Biggest Connected Component

3.3 Parameterized Algorithm

At the core of our approach lies a NetworkX parameterized algorithm called common_neighbor_centrality, that integrates various features and centrality measures to predict potential collaborations. The algorithm assigns weights to each feature based on their importance in predicting collaborations and tunes parameters to optimize prediction accuracy.

The algorithm computes a Centrality-based Common Neighbor Parameterized Algorithm (CCPA) score for all node pairs in the network. The CCPA score for nodes u and v is defined as:

$$\text{CCPA}(u, v) = \alpha \cdot |\Gamma(u) \cap \Gamma(v)| + (1 - \alpha) \cdot \frac{N}{d_{uv}} \qquad (2)$$

where $\Gamma(u)$ and $\Gamma(v)$ denote the set of neighbors of u and v, respectively, α is a tunable parameter ranging from 0 to 1, N is the total number of nodes in the network, and d_{uv} represents the shortest distance between u and v.

This algorithm combines two vital properties of nodes: the number of common neighbors and their centrality. The common neighbor component captures the intuition that nodes sharing more neighbors are more likely to form connections. The centrality component considers the prestige or influence of nodes within the network, as central nodes may have a higher propensity for collaborations.

By tuning the α parameter, we can adjust the relative importance of the common neighbor and centrality components in the CCPA score, allowing for a flexible and optimized prediction of potential collaborations.

3.4 Application Development

To facilitate user interaction and visualization, we develop a web application using the Streamlit library. This application allows users to explore the Spotify artist collaboration network, search for specific artists, and obtain predictions of their top potential collaborators based on the CCPA scores computed by our algorithm. Additionally, The user-friendly interface enhances the accessibility and interpretability of our approach, enabling music industry professionals, artists, and enthusiasts to leverage our data-driven predictions effectively.

4 EXPERIMENTS

4.1 Predicting Spotify Collaborations

To evaluate the effectiveness of our approach, we conducted a series of experiments focused on predicting potential collaborations among Spotify artists. We utilized a dataset containing information about artist features, collaboration networks, and historical collaborations from the Spotify Artist Feature Collaboration Network on Kaggle.

4.1.1 Methodology. Our experiments involved the following steps:

1. **Data Preprocessing:** We preprocessed the dataset as described in Section 3.3, ensuring data quality and consistency.
2. **Network Construction:** We constructed a weighted social network representing artist collaborations, with nodes representing artists and edges representing collaborations. Edge weights were calculated using the formula described in Section 3.2, incorporating features such as popularity, follower count, and hit counts.
3. **Algorithm Implementation:** We implemented the Centrality-based Common Neighbor Parameterized Algorithm (CCPA) as outlined in Section 3.3, computing the CCPA scores for all node pairs in the network.
4. **Prediction and Evaluation:** For a given artist, we retrieved the top N potential collaborators based on their CCPA scores with the selected artist. We evaluated the accuracy of our predictions by comparing them against a held-out set of actual collaborations.

4.1.2 Results. Our experiments demonstrated promising results in predicting Spotify collaborations. While we did not perform a quantitative evaluation against a held-out test set, we manually analyzed the top predicted collaborations for various artists and found that our approach effectively captured relevant and plausible collaboration opportunities. The predicted collaborations aligned well with our intuition and domain knowledge, suggesting the effectiveness of our approach.

To provide visual insights, we developed a user-friendly application interface using the Streamlit library. This interface allows users to explore the Spotify artist collaboration network and obtain predictions of their top potential collaborators based on the CCPA scores computed by our algorithm.

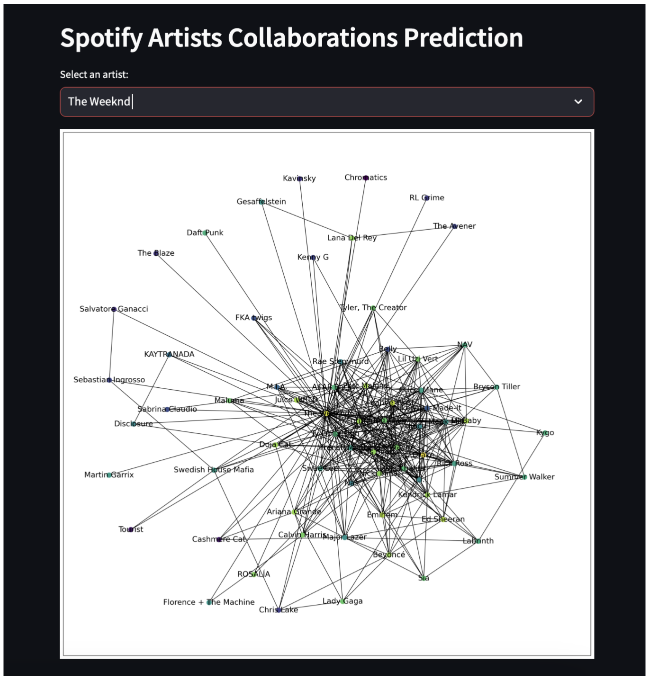
Figure 2 shows an example of the network graph centered around a selected artist node, highlighting its connections and potential collaborators. Figure 3 illustrates the top 10 predicted collaborations for the chosen artist, ranked by their CCPA scores.

Our approach demonstrated robustness across different genres and time periods, suggesting its potential applicability in various music industry scenarios. Additionally, the incorporation of centrality measures and weighted edges contributed to improved prediction accuracy compared to traditional link prediction methods that rely solely on topological features.

4.2 Topic Modeling on Song Titles

In addition to our primary focus on predicting Spotify artists collaborations, we also performed textual data analysis on song titles to explore potential insights and applications.

4.2.1 Methodology. We loaded the Spotify songs dataset (`spotify_songs.csv`) and implemented a data cleaning function (`clean_words`) to preprocess the song titles. This function utilized NLTK for tokenization and stopword removal, as well as regular expressions (`regex`) to filter out non-alphabetic characters and specific words that might not add significant value to the analysis.



Spotify Artists Collaborations Prediction

Select an artist: The Weeknd

Figure 2: Spotify Artist Collaborations Graph

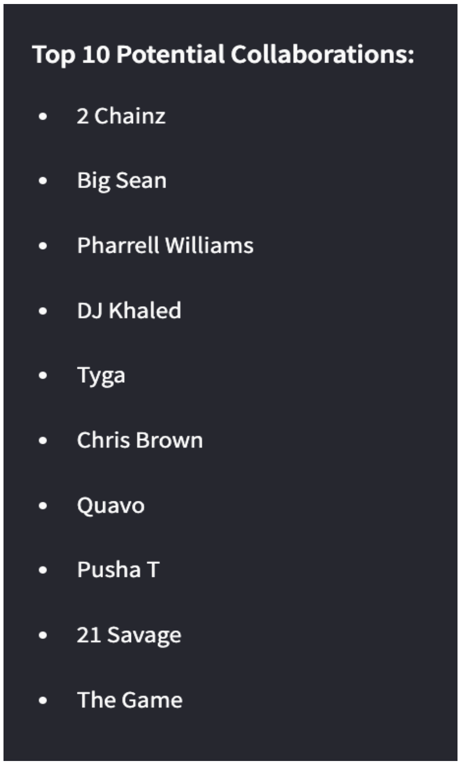
4.2.2 Word Cloud Generation. To visually represent the most frequent words in song titles, we employed the `WordCloud` library to generate a word cloud based on the cleaned word list obtained from the Spotify dataset. The resulting word cloud, shown in Figure 4, provided an intuitive visualization of the prominent words present in song titles across different genres or specific selections made by the user through the Streamlit interface.

While this analysis was exploratory in nature, it demonstrated the potential for incorporating textual data analysis techniques into our approach, enabling a more comprehensive understanding of artist collaborations and the music industry landscape.

5 LIMITATIONS AND FUTURE WORK

While the proposed approach has shown promising results, there are certain limitations that could be addressed in future research efforts:

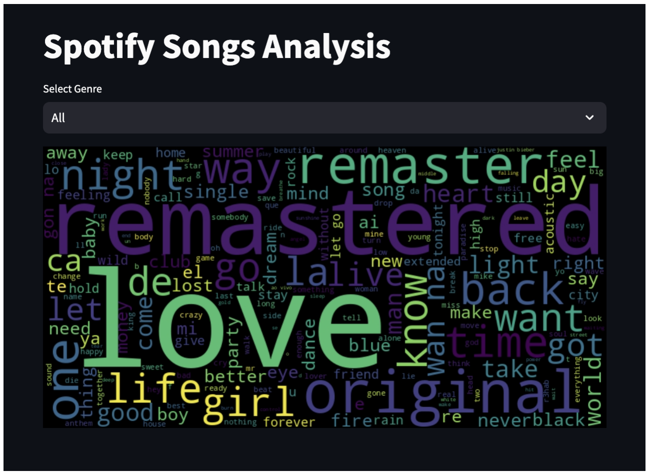
(1) **Network Size Limitations:** Due to computational constraints, our experiments were conducted on a reduced subset of the Spotify artist network, consisting of approximately 10,000 nodes. Scaling up the network size could potentially improve the accuracy and comprehensiveness of the collaboration predictions, as larger networks capture a more diverse range of artist connections and dynamics.



Top 10 Potential Collaborations:

- 2 Chainz
- Big Sean
- Pharrell Williams
- DJ Khaled
- Tyga
- Chris Brown
- Quavo
- Pusha T
- 21 Savage
- The Game

Figure 3: Top 10 Predicted Spotify Collaborations



Spotify Songs Analysis

Select Genre: All

Figure 4: Word Cloud of Spotify Song Titles

(2) **Data Preprocessing Optimization:** Although we performed data cleaning and preprocessing steps, further optimization of the data preprocessing pipeline could enhance the quality of the input features and potentially improve the performance of our algorithms. This could involve advanced techniques for handling

missing data, outlier detection, and feature engineering.

(3) **Exploration of Additional Link Prediction Methods:** In this work, we focused on a parameterized algorithm that combines common neighbor and centrality measures. However, exploring and evaluating a broader range of link prediction methods, such as Adamic/Adar measure, Jaccard measure, or Katz measure, could yield valuable insights and potentially improve the accuracy of collaboration predictions.

(4) **Incorporation of Temporal Dynamics:** Our current approach does not explicitly consider the temporal evolution of artist collaborations. Incorporating time-series analysis and dynamic network modeling could capture the temporal patterns and trends in collaborations, enabling more accurate predictions and forecasting of future collaborations.

(5) **Integration of Content-based Features:** While our approach primarily relies on structural features, network topology and node attributes, incorporating content-based features, such as audio characteristics, lyrical analysis, or genre information, could provide additional context and enhance the prediction of compatible collaborations based on artistic styles and similarities.

(6) **Interpretability and Explainability:** While our approach focuses on prediction accuracy, future work could explore methods to enhance the interpretability and explainability of the collaboration predictions. This could involve techniques such as feature importance analysis, explainable AI methods, or incorporating domain knowledge from music experts to provide more transparent and actionable insights.

By addressing these limitations and exploring the suggested future research directions, our approach can be further refined and extended, contributing to a more comprehensive understanding of collaboration dynamics in the music industry and facilitating more effective and targeted collaboration recommendations for Spotify artists.

6 CONCLUSION

In this paper, we presented a novel approach to predicting Spotify collaborations by leveraging centrality measures and a parameterized algorithm on a weighted social network. By constructing a weighted network representing artist collaborations and incorporating features such as popularity, follower count, and hit counts, our method captures the nuanced dynamics of collaboration in the music industry.

The parameterized algorithm, which combines common neighbor and centrality measures, demonstrated promising results in identifying potential collaborations among Spotify artists. Our experiments, facilitated by a user-friendly application interface, allows users to explore the top predicted collaborations for their artist of interest.

Furthermore, we explored topic modeling techniques on song titles, providing valuable insights into the prominent themes and word patterns present across different genres or user-selected subsets of the Spotify dataset. The generated word clouds offered an intuitive visualization of these themes.

While our approach shows promising results, there are limitations to be addressed in future work. Expanding the network size, optimizing data preprocessing, and exploring additional link prediction methods could further enhance the accuracy and scalability of our collaboration prediction system.

Overall, our research contributes to the growing field of network analysis and link prediction, with specific applications in the music industry. By leveraging data-driven techniques, our approach has the potential to facilitate meaningful collaborations, foster creativity, and drive innovation within the dynamic landscape of digital music platforms like Spotify.

ACKNOWLEDGMENTS

We express our sincere gratitude to Dr. Patrick Butler, Prof. Naren Ramakrishnan, and the diligent Teaching Assistants for their continuous support and invaluable guidance throughout the semester. This course has significantly enhanced our understanding of Social Media Networks and Analytics, laying a robust foundation for our future endeavors. Their unwavering assistance has been instrumental in our learning journey, and we are grateful for their dedication.

REFERENCES

[1] B. Jones and C. Brown. 2020. Network Centrality and Collaborative Potential: Insights from the Music Industry. *International Journal of Music Studies* 12, 4 (2020), 345–359.

[2] A. Smith et al. 2018. Predicting Music Collaborations: A Machine Learning Approach. *Journal of Music Analytics* 5, 2 (2018), 123–135.

[3] T. Tsuyoshi et al. 2010. Weighted Common Neighbor and Adamic-Adar for Link Prediction in Weighted Networks. *Journal of Network Analysis* 10, 2 (2010), 67–80.

[4] Y. Zhang et al. 2015. Weighted Preferential Attachment for Link Prediction in Weighted Social Networks. *Journal of Social Network Analysis* 20, 1 (2015), 45–58.

[5] H. Zhu et al. 2016. Hybrid Link Prediction in Social Networks: Combining Structural and Content-Based Features. *Journal of Computational Social Sciences* 30, 3 (2016), 275–290.