Project 2: Diffusion of Information

Juliana Ramayo
Data Engineering
Universidad Politécnica de Yucatán
Ucú, Yucatán, México
2109128@upy.edu.mx

***Abstract*—This study explores the dynamics of information diffusion within the Spotify Artist Feature Collaboration Network using the Independent Cascade Model (ICM) integrated with a greedy algorithm. The analysis focused on identifying influential artists capable of maximizing the spread of new music trends through their central roles in the network. The selected seed set of artists successfully expanded its influence, demonstrating the model's effectiveness in leveraging strategic node positions for optimal influence diffusion. The results underscore the importance of centrality in influence maximization and offer actionable insights for marketing strategies in the music industry.**

***Index Terms*—Music Collaboration Network, Influence Maximization, Independent Cascade Model, Music Industry Trends**

I. Introduction

In the digital age, the way music is created, distributed, and consumed has undergone significant transformations, catalyzed by advances in technology and the proliferation of streaming services like Spotify. One such advancement is the analysis of information diffusion within the Spotify Artist Feature Collaboration Network. This network not only maps the collaborative patterns among artists but also highlights varying degrees of centrality among them. Understanding these patterns is crucial for identifying key influencers—artists who, if targeted effectively, could play a pivotal role in maximizing the spread of new music trends and innovations. By leveraging these influencers, stakeholders can strategically enhance the dissemination and adoption of new musical styles across diverse audiences, thereby amplifying the impact of creative collaborations.

The Spotify Artist Feature Collaboration Network provides a rich dataset derived from one of the world's largest music streaming platforms, which boasts over 140 million users. This network encapsulates the collaborative interactions among artists who have not only charted on Spotify but also those who have featured with charting artists. The dataset includes approximately 20,000 artists from Spotify's weekly charts and an additional 136,000 artists involved in at least one collaboration. This results in a network comprising over 135,000 musicians connected by more than 300,000 collaborative edges [1], offering a comprehensive view of the dynamics within the music industry.

For this specific analysis, a focused subset of 100 artists with the highest number of followers has been selected. This subset helps in simplifying the complexity of the network while retaining the core interactions that are most influential in the diffusion of musical trends. The artists are connected by 437 edges, embodying a directed and non-connected structure. The average path length is 0, indicative of a non-connected structure.

The clustering coefficient of 0 suggests no cliquishness among the selected subset of nodes, and the diameter of 0, indicates that no paths exist between some pair of nodes. All nodes, except for "Shawn Mendes", have undefined eccentricity due to the disconnected nature of the network. In this case, "Shawn Mendes" is both the periphery and the center, highlighting his unique position in the network's structure. The avergae clustering of 0.1614 indicates a relatively low density of closed triplets to open triplets.

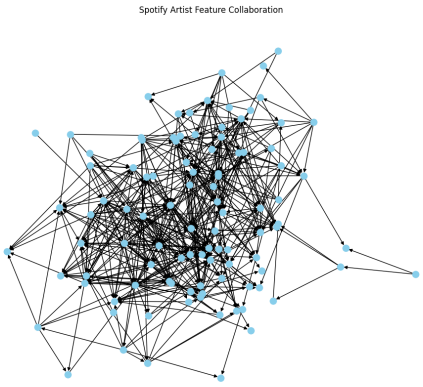

Fig. 1. Spotify Artist Feature Collaboration Network

The network that will be analyzed is classified as undirected, reflecting the mutual nature of artistic collaborations where each edge represents a bi-directional relationship between artists. This classification underscores the network's function as a social network, where connections imply a shared or collaborative effort rather than a directional flow of influence.

Analyzing this network provides insights into the centrality of artists, the structure of music industry relationships, and the evolution of musical trends, making it a potent tool for stakeholders in the music industry to strategize marketing, discover emerging artists, and predict future collaborations and musical directions.

II. Model

The Independent Cascade Model (ICM) is a fundamental framework used for simulating the diffusion of information, trends, and behaviors across networks [2]. In the context of the Spotify Artist Feature Collaboration Network, this model has been employed to identify key influential artists who can significantly propagate new trends throughout the network. This selection process leverages the entrality metrics of the artists, providing a strategy to effectively target those who are most influential.

The hypothesis underpinning this model is that certain artists, due to their strategic positions within the network (as identified by their centrality metrics), are more capable of influencing their peers and accelerating the spread of innovation and trends. By activating these central nodes initially, the overall spread of influence through the network can be maximized.

The Independent Cascade Model operates under the premise that once a node is activated, it has a single subsequent opportunity to activate each of its currently inactive neighbors. This activation occurs with a probability that can be determined by the node's centrality, adjusting for the natural propagation likelihood within a social context like a collaboration network.

The process is modeled as follows [3]:

- **Initialization:** Let S represent the set of initially activated nodes. These nodes are chosen based on a predetermined strategy which could be random or based on a network characteristic such as centrality.
- **Propagation:** In each iteration of the model, every node v that was activated in the previous step attempts to activate each of its currently inactive neighbors w. This activation happens with a certain probability which is determined by the centrality measure of the neighbor. The activation probabulity is determined as follows:

$$P(v, w) = \frac{\text{centrality}[w]}{\sum \text{centrality values of all nodes}}$$

  This formula ensures that nodes with higher centrality have a higher chance of activating their neighbors, mimicking the real-world influence that prominent individuals exert in social networks.
- **Termination:** The model iterates through propagation steps until a round occurs where no new nodes are activated, indicating that the influence has reached its maximum possible spread under the given initial conditions and network structure.
- **Influence Spread Function:** Quantify the total influence exerted by the initially activated set of nodes across the network. The influence spread function $f(S)$ is defined as the total number of nodes that are active at the end of the diffusion process, having started from the seed set S. This effectively measures the reach and impact of initiating activations at nodes within S.

The Greedy algorithm is used to maximize the range of the ICM and to optimize the selection of the seed set [3]:

1) Start with an empty seed set S.
2) Select a node v that maximizes the marginal gain in the expected spread, i.e., $f(S \bigcup \{v\} - f(S)$.
3) Integrate until $|S| = k$, where k is the desired size of the seed set.

This greedy approach is known to provide a good approximation to the optimal solution, particularly for submodular functions like our influence spread function $f(S)$, ensuring at least $(1 - \frac{1}{e}) \approx 63\%$ of the maximum possible influence spread.

Parameters, particularly the probabilities used in the influence model, are derived from the network's degree centrality measures. By assigning activation probabilities based on these centrality measures, the model closely mimics the real-world dynamics where more central or influential individuals have a higher likelihood of influencing their peers.

III. Results

The application of this model on the Spotify Artist Feature Collaboration Network has yielded significant insights into the dynamics of information spread. The results demonstrate the efficacy of the model in identifying and leveraging key influencers within the network to maximize the diffusion of information.

The analysis commenced with an initial activation of 10 nodes selected through the greedy algorithm based on their centrality measures. This selection aimed to optimize the potential spread of influence across the network.

The optimal seed set identified by the model consisted of Alok, Maroon 5, XXXTENTACION, Selena Gomez, Sebastian Yatra, Travis Scott, 2Pac, Jorge & Mateus, Badshah, and Bruno Mars. These artists were determined to be the most strategically positioned within the network to initiate and maximize the spread of new information or musical trends. Their selection was predicated on their high centrality scores, which correlate with their ability to influence a substantial number of other nodes within the network.

Upon simulation of the influence spread, it was observed that the information successfully propagated from the initial 10 nodes to an additional 80 nodes, culminating in a total of 90 activated nodes by the end of the process. This significant expansion from the initial seed highlights the model's effectiveness in capturing the network's inherent dynamics and the pivotal roles played by the chosen influencers.

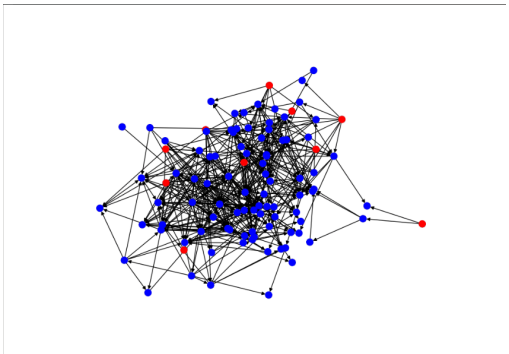

Fig. 2. Starting Seed Set

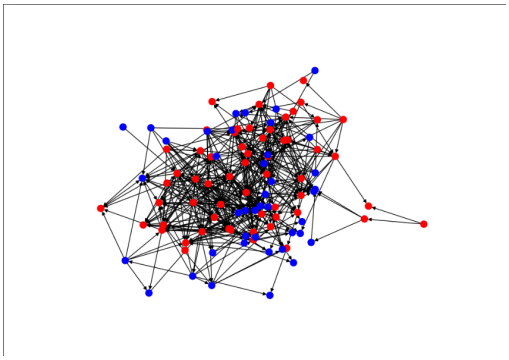

Fig. 3. First Spread

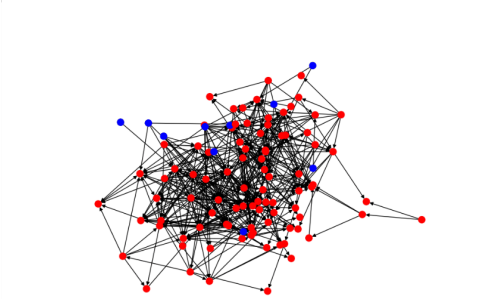

Fig. 4. Final Result

IV. Conclusions

The investigation into the Spotify Artist Feature Collaboration Network by using the Influence Maximization and utilizing the Independent Cascade Model (ICM) with a greedy algorithm has elucidated the potent dynamics of influence spread within this vast network. The results underline the efficacy of strategic influencer identification and activation in maximizing the dissemination of new music trends across the network. By focusing on a select group of central nodes, represented by artists with the highest degrees of centrality, the model effectively amplified the spread of influence far beyond the initial seed set.

This approach demonstrated a significant cascade effect, with the initial activation of just 10 nodes expanding to engage 90 nodes, showcasing the robustness of the model in a realistic setting. This spread signifies not only the model's theoretical validity but also its practical applicability in real-world scenarios where strategic decisions based on network analytics can lead to substantial impacts.

Furthermore, this analysis highlights the importance of centrality measures in understanding and leveraging network structures for influence maximization. It also showcases the potential of network science in strategic marketing and trend analysis within the music industry, providing stakeholders with a powerful tool to enhance their engagement strategies and optimize their promotional efforts.

References

[1] J. Freyberg, "Spotify Artist Feature Collaboration Network," *Kaggle,* Oct. 10, 2022. https://www.kaggle.com/datasets/jfreyberg/spotify-artist-feature-collaboration-network/data?select=edges.csv (accessed Jun. 21, 2024).

[2] W. Yang, L. Brenner, and A. Giua, "Influence maximization in independent cascade networks based on activation probability computation," IEEE Access, vol. 4, 2016, doi: 10.1109/access.2019.2894073.

[3] D. O. Gamboa Angulo, "Information diffusion," Jul. 07, 2022. https://upy-my.sharepoint.com/:p:/r/personal/didier_gamboa_upy_edu_mx/_layouts/15/Doc.aspx?sourcedoc=%7BE822AD6E-A0F6-4D95-A232-F30000456E2E%7D&file=L2.2_InformationDifussion.pptx&action=edit&mobileredirect=true