Behavioral and Acoustic Artist Graphs for Music Community Detection and Genre Inference

Manuel Bezerra-Brandao[1], **Lilian Berton**[1]

[1]Instituto de Ciência e Tecnologia – Universidade Federal de São Paulo (ICT-Unifesp)
São José dos Campos – SP – Brazil

{mbrandao25,lberton}@unifesp.br

Abstract. *This work compares behavioral and acoustic artist graphs for community detection and missing-genre inference. A behavioral graph is constructed from repeated artist proximity in ten Last.fm listening histories, while an acoustic graph is constructed from aggregated AcousticBrainz low-level descriptors. Genre labels, tags, textual metadata, and high-level acoustic annotations are excluded from graph construction; MusicBrainz genres are used for evaluation and downstream inference. Louvain, Leiden, Node2Vec, Graph Autoencoders, and propagation controls are evaluated. The observed behavioral results show higher adjusted genre alignment and Macro-F1; however, differences in artist coverage, genre spaces, and class distributions prevent direct comparison between views. Acoustic similarity instead provides a complementary, coverage-limited view of artist relatedness. In a stratified manual audit, considering the top three behavioral Node2Vec suggestions provided at least one genre judged useful for 90.7% of reviewed artists. These findings support graph-based candidate generation for human-assisted metadata curation.*

1. Introduction

Community detection can reveal latent graph structure by identifying densely connected groups [Fortunato 2010, Bedi and Sharma 2016, Zheng et al. 2021]. In music analysis, artist graphs support recommendation, catalog exploration, metadata enrichment, and visualization [Vlegels and Lievens 2017]. However, the recovered structure depends strongly on how artist relationships are defined.

This work compares two complementary notions of artist relatedness. The behavioral view connects artists that repeatedly occur near each other in Last.fm listening sessions [Haupt 2009], capturing co-listening patterns derived solely from observed sequences. The acoustic view connects artists with similar aggregated AcousticBrainz low-level descriptors [Porter et al. 2015], capturing timbral, rhythmic, spectral, and tonal similarity. Artists may therefore share listening contexts without similar acoustic profiles, while acoustically similar artists may belong to different cultural or genre contexts.

MusicBrainz genres [Stutzbach 2011] serve as an external reference for evaluation and downstream inference. For each labeled artist, the normalized annotation with the highest MusicBrainz count is selected as its dominant genre. Community partitions are evaluated against these labels, while artists without usable annotations remain in the graphs for missing-genre suggestion. Genre labels, tags, textual metadata, and high-level acoustic annotations are excluded from graph construction.

The contributions are threefold: (i) an artist-level operationalization of behavioral and low-level acoustic relatedness without semantic labels in graph construction; (ii) a multi-method evaluation using Louvain [Blondel et al. 2008], Leiden [Traag et al. 2019], Node2Vec [Grover and Leskovec 2016], GAE [Kipf and Welling 2016], and propagation controls to characterize community alignment, partition stability, and genre-inference performance within each coverage-specific graph view; and (iii) a three-reviewer manual audit supporting graph-based suggestions as candidates for human-assisted metadata curation.

2. Related Work

Graph-based music analysis has modeled artist or song relationships through different forms of similarity. Content-based approaches connect musical items using acoustic or rhythmic descriptors. Prior work on song networks built from rhythmic similarity between percussion tracks showed that audio-based relationships can recover community structures associated with known genres [Correa and Levada 2011].

Behavioral and social approaches instead derive relationships from user activity, tags, or listening preferences. Artist–tag networks derived from individual listening data have been analyzed with Social Network Analysis and Louvain community detection to identify thematic substructures in personal music collections [Dolgikh and Jelínek 2015]. More broadly, these studies show that the musical meaning of detected communities depends strongly on how graph edges are defined [Vlegels and Lievens 2017].

Existing studies generally examine content-based and behavioral music networks separately, often using different nodes, labels, and evaluation criteria. Evidence on how these edge definitions affect artist communities and missing-metadata inference within a shared methodological framework therefore remains limited. This work addresses that gap as a comparative case study, focusing on community stability, external genre alignment, interpretability, and downstream suggestion.

3. Methodology

The configurations reported in this section were selected from preliminary sensitivity analyses, with the full sensitivity results and implementation details made available in the accompanying repository.

3.1. Data Sources and Preprocessing

This study combines three complementary data sources. Last.fm listening histories provide timestamped listening events, referred to as *scrobbles*, used as behavioral evidence of artist proximity [Haupt 2009]. MusicBrainz provides standardized artist identifiers and external genre annotations [Stutzbach 2011]. AcousticBrainz provides low-level audio descriptors for MusicBrainz recordings, enabling an acoustic comparison view [Porter et al. 2015].

The behavioral dataset comprises ten pseudonymized Last.fm listening histories from highly active users in Peru and Brazil, selected as a focused case study of contemporary and international popular-music listeners. Participants were informed about the study and agreed to the use of their data, which were publicly accessible at collection

time and exported using the LastFM Data Export tool[1]. Usernames were replaced with study identifiers, and raw histories are not released. Although detailed listening patterns retain some re-identification risk, user-level results are reported only in pseudonymized aggregate form.

After exact-duplicate removal and artist-name normalization, the merged dataset contains 1,760,085 scrobbles from 22,596 artists, spanning July 10, 2006 to June 10, 2026. Available artist MBIDs were retained; missing MBIDs were assigned only when a normalized artist name uniquely matched a previously resolved identifier. Artists without recoverable MBIDs were preserved using stable name-derived identifiers. Figure 1 summarizes yearly activity and the user-level composition of the thirty most-listened artists.

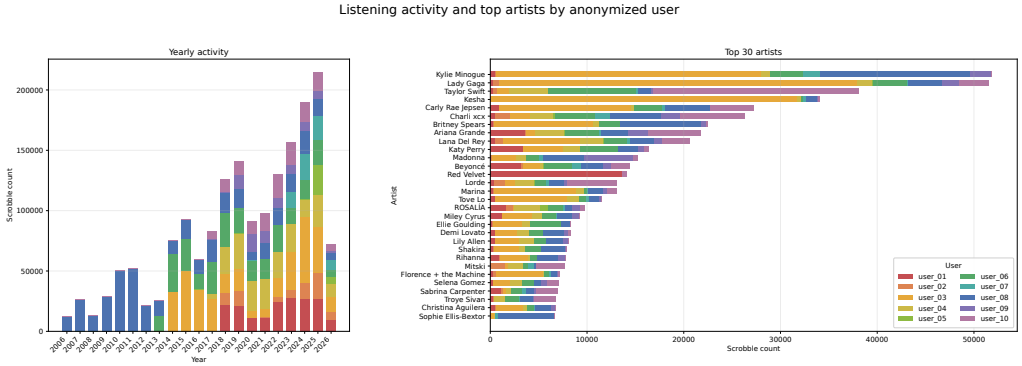


Figure 1. Yearly listening activity by anonymized user and user-level composition of the thirty most-listened artists in the merged dataset.

MusicBrainz genre annotations are used as external labels for evaluation and supervised genre inference. For each labeled artist, a single dominant genre was selected as the MusicBrainz genre with the highest annotation count after normalization. In the artist set retained for the experiments, 3,493 artists have MusicBrainz genre annotations, 2,828 have no resolved MBID, and 1,834 have an MBID but no available genre annotation. Since MusicBrainz genres are collaboratively maintained and may be incomplete or noisy, they are treated as external reference labels rather than definitive ground truth.

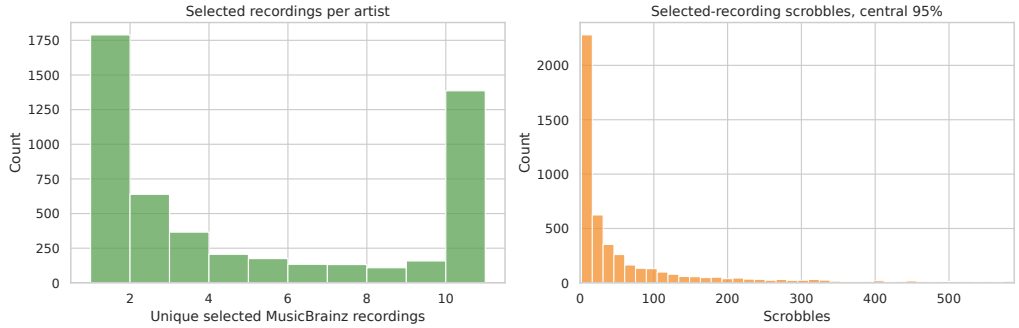


Figure 2. Coverage of selected MusicBrainz recordings with AcousticBrainz low-level descriptors.

AcousticBrainz data were collected for MusicBrainz recordings associated with the listened artists. For each artist, the pipeline considered the most-listened recordings

[1]Developed by `ghan64` and available at: `https://lastfm.ghan.nl/export/`

with available MusicBrainz recording identifiers and retrieved AcousticBrainz low-level data JSON files. High-level AcousticBrainz genre or mood annotations were intentionally excluded to avoid introducing genre information into the acoustic representation. In the current dataset, 1,672 selected recordings have AcousticBrainz descriptors, while 21,979 queried recordings were not found. The retained AcousticBrainz feature set contains 134 numerical variables: two numeric audio properties (track length and replay gain), two low-level scalar descriptors, 27 Bark-band means, 40 ERB-band means, 13 MFCC means, 9 spectral descriptors, 3 rhythm descriptors, 2 tonal scalar descriptors, and 36 HPCP means. Textual or semantic metadata fields, including artist names, album names, tags, and genres, were excluded. Figure 2 summarizes the selected-recording coverage and scrobble distribution.

3.2. Construction of the Artist Graphs

Two weighted artist graphs were constructed: a behavioral graph, which captures local proximity in listening sequences, and an acoustic graph, which captures similarity between aggregated low-level audio descriptors. Both graphs use artists as nodes, but their edges encode different notions of relatedness.

For the behavioral graph, scrobbles were first segmented into listening sessions using a 60-minute inactivity threshold. Within each session, candidate artist pairs were generated using a sequential window of five scrobbles: each artist was paired only with artists appearing in the next five positions of the same session, and self-pairs were discarded. Edge weights accumulated this evidence across sessions and users, with a distance-decay factor giving more importance to artists heard closer together in sequence. The accumulated weights were then normalized with a cosine-like activity correction to reduce the dominance of very frequently heard artists. Artists with fewer than three scrobbles were excluded, and edges were retained only when supported by at least two distinct sessions.

For the acoustic graph, recording-level AcousticBrainz descriptors were first aggregated into artist-level vectors using up to ten selected recordings per artist. The resulting 134-dimensional artist vectors were standardized, and pairwise acoustic similarities were computed. Since this similarity space is dense by construction, the graph was sparsified by retaining each artist's 20 nearest neighbors and applying a minimum similarity threshold of 0.40. High-level AcousticBrainz genre or mood annotations were not used.

Table 1 summarizes the final largest connected components used in the experiments. The behavioral graph is substantially larger than the acoustic graph because listening histories provide broader artist coverage than the subset of artists with matched AcousticBrainz descriptors. This coverage gap is considered later when interpreting both community detection and downstream genre-inference results.

Table 1. Summary of the artist graphs

Graph	Nodes	Edges	Main construction rule
Behavioral	8,155	216,806	Sequential window $k = 5$ and support ≥ 2 sessions
Acoustic	1,022	13,729	kNN $= 20$ and similarity ≥ 0.40

3.3. Community Detection and Representation Learning

The two artist graphs were evaluated with modularity-based methods, learned graph embeddings, and non-trainable propagation controls. Louvain and Leiden were

used as community-detection baselines [Blondel et al. 2008, Traag et al. 2019]. Node2Vec learned structural embeddings from biased random walks [Grover and Leskovec 2016], while Graph Autoencoders (GAE) provided a trainable alternative [Kipf and Welling 2016]. The propagation controls were included to distinguish graph smoothing from learned representations. Table 2 summarizes the final configurations.

For the propagation controls, X contained four log-transformed node statistics: scrobble count, session count, degree, and weighted degree. In the acoustic view, session count is zero, and the 134 low-level descriptors are used to construct edges but are not included directly in X. From the complete weighted adjacency A, the matrices were defined as $\tilde{A} = A + I$, $\tilde{D}_{ii} = \sum_j \tilde{A}_{ij}$, and $\hat{A} = \tilde{D}^{-1/2}\tilde{A}\tilde{D}^{-1/2}$. The evaluated variants were the unsmoothed statistics X, one-step smoothing $\hat{A}X$, and two-step smoothing $\hat{A}^2X$, each subsequently standardized and row-normalized. GAE instead used a separate five-feature matrix that additionally included user count and applied the same normalization to an unweighted training adjacency after holding out one tenth of the edges for validation. Session and user counts are zero in the acoustic GAE input. Representation-based methods were clustered with K-Means using $k = 54$. All stochastic methods were executed with seeds 42–46.

Table 2. Configuration of the evaluated methods

Method	Configuration
Louvain / Leiden	Weighted graphs; resolution $= 2.25$
Node2Vec	64 dimensions; 10 walks per node; walk length 40; window 5; 30 epochs; behavioral $p = 2.0$, $q = 1.0$; acoustic $p = 2.0$, $q = 2.0$
GAE	Two-layer encoder; hidden 128; dropout 0.3; learning rate 0.005; weight decay 0.0001; 100 epochs; latent dimensions 16 behavioral, 32 acoustic
Propagation controls	Unsmoothed node statistics X, one-step smoothing $\hat{A}X$, and two-step smoothing $\hat{A}^2X$
K-Means	For representation-based methods with $k = 54$ and 20 centroid initializations

3.4. Evaluation and Genre Inference

Community detection and clustering were evaluated on artists with available MusicBrainz dominant genres, used as external reference labels. Agreement was measured using Normalized Mutual Information (NMI), Adjusted Mutual Information (AMI), Adjusted Rand Index (ARI), and purity, while stability across stochastic runs was measured through average pairwise NMI and ARI. Because NMI and purity are sensitive to the number and size of clusters, interpretation prioritizes adjusted metrics and stability.

Genre inference was formulated as multiclass classification over MusicBrainz dominant genres. For each view, raw, smoothed, Node2Vec, and GAE representations were evaluated using class-weighted Logistic Regression. Genres with fewer than five labeled artists were excluded, and C was selected through inner cross-validation within each outer fold. Representations were learned transductively from the complete graph without using held-out genre labels. Because AcousticBrainz covered fewer artists, the two benchmarks differed in artist coverage and retained genres and are therefore not directly comparable. Table 3 summarizes the evaluation setup.

Table 3. Genre inference evaluation setup

Component	Configuration
Target labels	MusicBrainz dominant genres
Genre filter	Minimum support of five labeled artists
Classifier	Logistic Regression with balanced class weights
Regularization grid	$C \in \{0.01, 0.1, 1.0, 10.0, 100.0\}$
Model selection	Inner cross-validation within each outer fold
Evaluation	Five seeds and five stratified outer folds
Metrics	Accuracy, Macro-F1, balanced accuracy, and top-3 accuracy

Following evaluation, the behavioral representation with the highest Macro-F1 was used to suggest genres for unlabeled artists. Suggestions were organized into heuristic confidence bands based on the top-class probability and its margin over the second-ranked genre; because probabilities were uncalibrated, these bands served only to prioritize manual review. The suggestions were assessed separately by three contemporary and international popular-music enthusiasts based in Peru and Brazil. Two of the three reviewers were external to behavioral dataset construction and model development. The audit was intended as a qualitative assessment for human-assisted metadata curation rather than as an independent expert evaluation.

4. Experimental Results

This section evaluates the two graph views through community detection and missing-genre inference. For community detection, the analysis assesses whether the detected partitions are stable across runs and whether artists grouped in the same communities tend to share the same external MusicBrainz dominant genre labels. The inference benchmark evaluates whether graph-derived representations can predict dominant genres for labeled artists and generate candidates for artists without usable genre annotations.

4.1. Quantitative Community Alignment

This subsection evaluates partition stability and consistency with external MusicBrainz dominant genre labels within each graph view. The behavioral evaluation contains 3,493 labeled artists across 341 genres, whereas the acoustic evaluation contains 735 artists across 143 genres. Because the views differ substantially in artist coverage, sample composition, and label space, their metric values are not directly comparable as a controlled head-to-head evaluation. Table 4 therefore characterizes each view under its available coverage.

Table 4. Community alignment and partition stability results

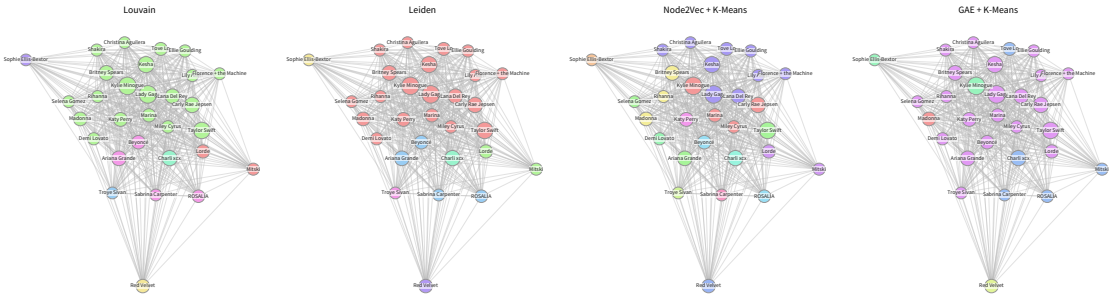
Graph	Method	NMI	AMI	ARI	Purity	Stab. NMI	Stab. ARI
Behavioral	Leiden	0.335	0.147	0.070	**0.223**	**0.877**	**0.852**
	Louvain	0.333	0.145	0.068	0.222	0.869	0.829
	Node2Vec	**0.345**	**0.151**	**0.075**	0.211	0.703	0.515
	GAE	0.277	0.081	0.037	0.169	0.682	0.443
Acoustic	Leiden	0.280	0.053	0.019	0.181	**0.736**	**0.628**
	Louvain	0.281	**0.055**	**0.020**	0.182	0.708	0.591
	Node2Vec	**0.432**	0.047	0.015	**0.233**	0.709	0.404
	GAE	0.421	0.040	0.016	0.220	0.705	0.423

Within the behavioral evaluation set, Node2Vec obtains the highest NMI, AMI, and ARI, indicating the strongest genre-related alignment among the behavioral methods. Leiden and Louvain remain close in AMI and ARI but are substantially more stable across runs, indicating more reproducible structural communities.

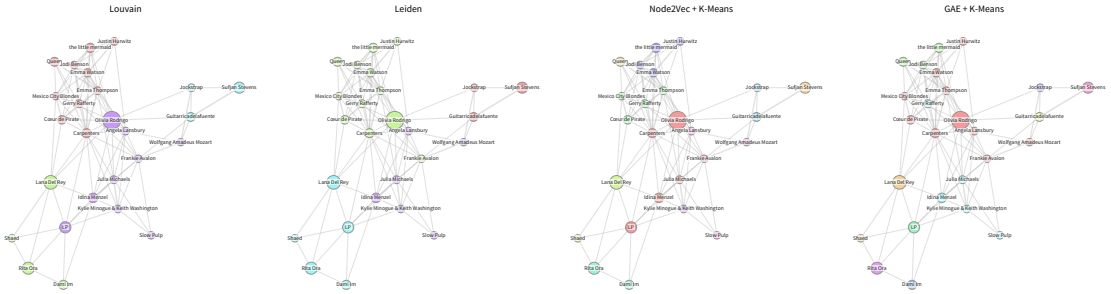
Within the acoustic evaluation set, Node2Vec and GAE obtain comparatively high NMI values, and Node2Vec has the highest purity, while their low AMI and ARI suggest that nominal agreement may partly reflect fragmentation or chance structure. Louvain achieves the highest adjusted scores among the acoustic methods, whereas Leiden provides the most stable partitions.

Overall, the behavioral view exhibits higher adjusted alignment under its broader available coverage, but the present design does not isolate graph-view effects from coverage and label-space differences. Within the behavioral graph, Node2Vec provides the strongest adjusted external alignment, while Louvain and Leiden prioritize stability. The acoustic graph remains a complementary, coverage-limited view.

4.2. Qualitative Community Characterization



(a) Behavioral graph



(b) Acoustic graph

Figure 3. Partition comparison across Louvain, Leiden, Node2Vec, and GAE on fixed local neighborhoods from the behavioral and acoustic graphs.

Figure 3 compares how the four methods partition view-specific representative neighborhoods. The behavioral view uses the induced subgraph of the 30 most-listened artists, whereas the smaller acoustic view uses a strongest-neighbor expansion from the most-listened artist with acoustic coverage. Nodes, edges, and layouts are fixed within each view, with only community colors varying across methods. Because the neighborhoods follow different selection procedures, the figure provides an illustrative rather than controlled cross-view comparison.

Table 5. Qualitative profiles of selected Leiden communities

ID	Preliminary qualitative profile
Behavioral graph	
0	Mainstream international pop and dance-pop, centered on high-activity pop divas and singer-songwriter-adjacent pop acts.
1	Pop, synth-pop, and alternative pop legacy neighborhood, mixing dance-pop, indie pop, and pop-rock references.
4	Contemporary pop and R&B crossover community, combining mainstream pop, vocal pop, and dance-pop/R&B-adjacent artists.
2	Alternative pop, indie-pop, and singer-songwriter crossover, connecting contemporary pop with older art-pop and vocal.
6	K-pop-centered community dominated by girl groups and related soloists.
3	Alternative pop, electropop, and experimental pop neighborhood, with strong club-pop and internet-pop connections.
7	Latin alternative, indie-pop, and contemporary crossover community, combining Spanish-language indie scenes with adjacent alternative pop.
5	Indie, alternative pop-rock, and singer-songwriter community, with Latin indie and anglophone indie-pop connections.
14	Contemporary anglophone pop and synth-pop community, centered on 2010s/2020s pop and internet-pop-adjacent artists.
9	Latin pop, Brazilian pop-funk, and club-oriented pop community.
12	Spanish-language synth-pop, electropop, and Latin indie community; *blackbear* appears as an outlier in this group.
8	Indie rock, synth-pop, and electro-indie community, mixing indie-pop bands with electronic-leaning alternative acts.
Acoustic graph	
6	Pop-oriented acoustic neighborhood, with dance-pop and contemporary vocal-pop textures.
8	Contemporary pop and R&B-adjacent acoustic neighborhood, mixing vocal pop, soft pop, and electronic/pop production traits.
9	Club-oriented pop, electronic, and hip-hop-adjacent acoustic neighborhood with energetic production similarities.
1	Dance-pop and synth-pop acoustic neighborhood.
3	High-energy pop and dance-pop acoustic neighborhood.
5	Mixed acoustic neighborhood, combining pop, theatrical recordings, classic vocal material, and softer singer-songwriter.
0	Bright pop and dance-pop acoustic neighborhood, with 1990s/2000s pop and dance-pop references.
7	Softer alternative-pop and dream-pop-adjacent acoustic neighborhood.
2	Folk, singer-songwriter, and organic vocal texture neighborhood.
4	Band-oriented pop and rock acoustic neighborhood.
13	Contemporary dance-pop and electropop acoustic neighborhood.
10	Smooth pop and R&B-adjacent acoustic neighborhood.

Note: Profiles are qualitative summaries of selected Leiden communities. Behavioral communities are visualized in Figure 4; acoustic community visualizations are available in the accompanying repository. Profiles are intended to support interpretation and should not be read as ground-truth genre labels.

Louvain and Leiden produce the most similar behavioral partitions, while Node2Vec changes more boundary assignments and GAE appears more diffuse. The acoustic neighborhood is less directly interpretable and is partitioned more variably by Node2Vec and GAE. Leiden was selected for subsequent characterization because it achieved the highest stability in both views while remaining competitive in external genre alignment, allowing the analysis to use a consistent structural reference.

Figure 4 displays the 10 most-listened artists in selected high-activity behavioral Leiden communities. Table 5 summarizes profiles for both views, with acoustic visualizations available in the repository. Behavioral profiles frequently reflect genre, scene, language, or cultural listening contexts, whereas acoustic communities may connect artists through shared acoustic or production traits without clear cultural boundaries. These profiles are post-hoc exploratory interpretations rather than blinded qualitative coding or ground-truth labels.

Overall, these exploratory observations are consistent with the quantitative alignment patterns: behavioral communities more often admit genre, scene, language, or cultural interpretations, whereas acoustic communities tend to reflect shared acoustic or production traits, although the view-specific neighborhood selection prevents a controlled qualitative comparison.

4.3. Genre Inference and Manual Audit

After genre filtering, the behavioral benchmark included 3,101 labeled artists across 109 retained genres, whereas the acoustic benchmark included 567 artists across 33 genres.

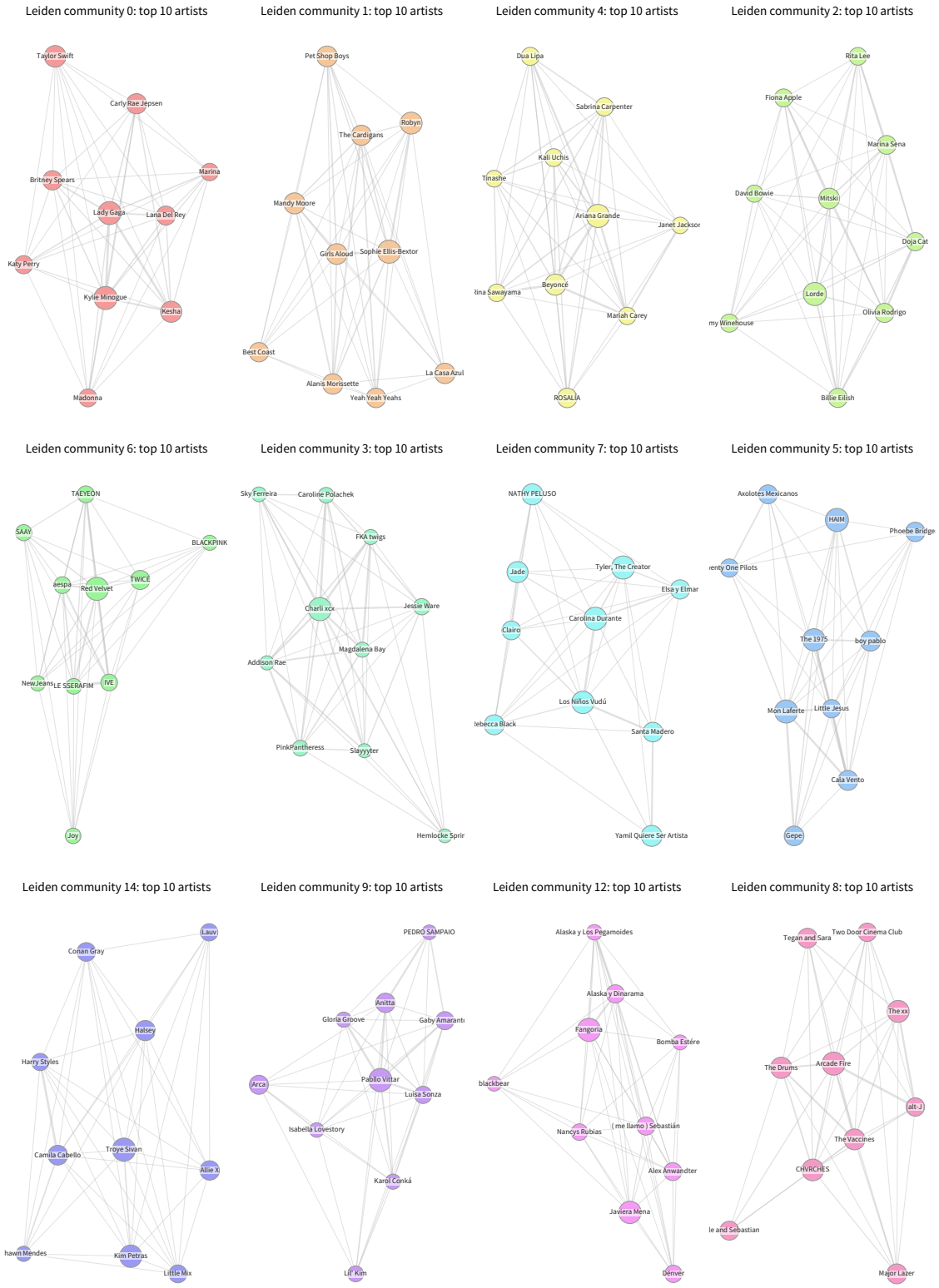


Figure 4. Selected high-activity Leiden communities in the behavioral graph.

Because these benchmarks differ in artist coverage, label space, and class distribution, their scores are not directly comparable; Table 6 should therefore be interpreted primarily within each graph view. In the behavioral benchmark, Node2Vec achieves the best model results across all reported metrics, including a macro-F1 of 0.099. In the acoustic benchmark, Node2Vec obtains the highest model accuracy, macro-F1, and top-3 accuracy, while GAE achieves the highest balanced accuracy. The acoustic dummy exceeds the learned representations in accuracy and top-3 accuracy, consistent with a strongly imbalanced label distribution.

Table 6. Genre inference performance by graph representation.

Graph	Representation	C	Accuracy	Balanced acc.	Macro-F1	Top-3 acc.
Behavioral	Node2Vec	0.1	**0.130**	**0.130**	**0.099**	**0.256**
	GAE	100.0	0.069	0.069	0.039	0.111
	Smooth-2	100.0	0.034	0.019	0.007	0.060
	Smooth-1	100.0	0.022	0.016	0.006	0.054
	Raw	100.0	0.016	0.014	0.005	0.040
	Dummy (most frequent)	–	0.113	0.009	0.002	0.222
Acoustic	Node2Vec	0.1	**0.101**	0.075	**0.069**	**0.279**
	GAE	100.0	0.049	**0.078**	0.045	0.148
	Smooth-2	10.0	0.026	0.044	0.023	0.089
	Smooth-1	100.0	0.017	0.038	0.017	0.066
	Raw	10.0	0.019	0.040	0.012	0.075
	Dummy (most frequent)	–	***0.199***	0.030	0.010	***0.376***

Note: Bold values indicate the best model result within each graph view. Bold italic values indicate dummy results that exceed all model representations for that metric. The C column reports the median Logistic Regression regularization value selected across evaluation runs.

Absolute macro-F1 remains modest, reflecting the long-tailed genre spaces and the evaluation of inherently multigenre artists against a single dominant label, which may penalize plausible alternatives. Improvements over the dummy baselines in macro-F1 and balanced accuracy nevertheless indicate limited genre-related signal within each benchmark. Because coverage and label spaces differ, the scores do not establish a general advantage of either view. The representations are therefore better suited to candidate generation for human-assisted curation than to automatic genre assignment.

For candidate generation, the behavioral view was selected because it provided broader coverage of labeled artists, with Node2Vec chosen for achieving the highest macro-F1 within that benchmark. Its suggestions were grouped into heuristic confidence bands based on the top-class probability and prediction margin; because probabilities were uncalibrated, the bands served only to prioritize manual review.

The audit included 50 artists from each confidence band and three reviewers working independently. For each artist, the top three predicted genres were assessed as *correct*, *plausible*, *unclear*, or *incorrect*. Reviewer judgments were aggregated by simple majority. When one judgment was *unclear* and the remaining substantive judgments disagreed, the more favorable substantive label was retained following the order *correct* $>$ *plausible* $>$ *incorrect*. Cases with three different substantive judgments were marked as *no consensus*. A prediction was considered exact when assessed as *correct* and useful when assessed as either *correct* or *plausible*.

Figure 5 shows that, within the stratified audit sample, the top-ranked suggestion was judged exact for 53.3% of artists and useful for 72.7%. Considering the top three suggestions increased these rates to 70.7% and 90.7%, respectively, with higher-confidence bands showing greater reliability. Because the audit sampled the same number of artists

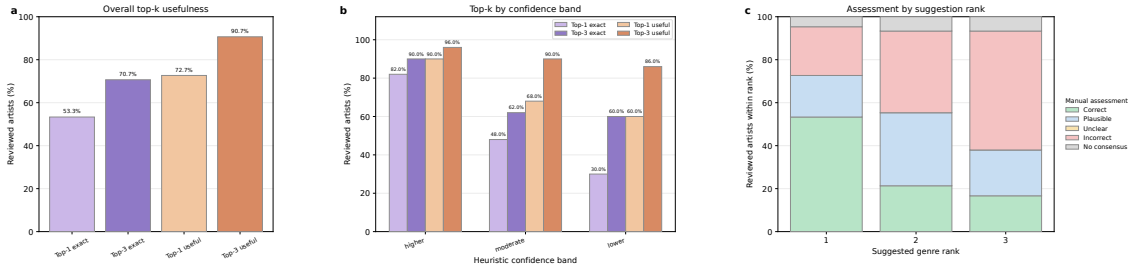


Figure 5. Manual audit of top-3 genre suggestions from the behavioral Node2Vec model.

from each band, these unweighted rates describe the reviewed sample rather than the full population of unlabeled artists. The audit also revealed culturally structured patterns: Brazilian artists associated with broad genres such as pop, rap, hip hop, or rock were often suggested as *mpb*, whereas Korean artists were frequently ranked as *k-pop* followed by *j-pop*. These patterns suggest that the behavioral representation may capture cultural and geographic listening contexts in addition to genre similarity. Overall, the results support its use for candidate generation in human-assisted metadata curation rather than fully automatic annotation.

5. Conclusion

This work compared behavioral and acoustic artist graphs constructed without genre labels, tags, textual metadata, or high-level acoustic annotations. The observed behavioral results showed stronger adjusted genre alignment and inference signal, although differences in coverage, label spaces, and class distributions prevent direct comparison and do not establish general superiority. Leiden produced the most stable partitions, while Node2Vec achieved the highest macro-F1 within the behavioral benchmark. In the stratified manual audit, considering the top three rather than only the top-ranked suggestion increased judged usefulness from 72.7% to 90.7% within the reviewed sample, supporting candidate generation for human-assisted metadata curation.

These findings constitute a focused case study based on ten highly active, pop-oriented listeners and should not be interpreted as population-level evidence. Further limitations include restricted AcousticBrainz coverage, incomplete single-label MusicBrainz references, transductive evaluation, uncalibrated confidence bands, and a favorable audit tie-breaking rule that may overestimate usefulness. Future work should evaluate broader listener samples and conduct paired cross-view comparisons using the same artists, genres, folds, and class distributions. Larger audits should employ inter-rater agreement measures, predefined neutral aggregation rules, and blinded paired assessments, alongside multilabel references, calibrated uncertainty, and multimodal graph approaches.

Code Availability

A repository containing the implementation, reproducibility artifacts, sensitivity analyses, and supplementary acoustic community visualizations is available at `https://github.com/mbezerrabrandao/music-graphs`. The supplementary visualizations correspond to the acoustic profiles summarized in Table 5.

Acknowledgments

The authors acknowledge the financial support provided by the São Paulo Research Foundation (FAPESP) under grant numbers 2024/22546-0 and 2020/09850-0. The authors also acknowledge the use of OpenAI's ChatGPT for language editing and writing support. All scientific content, experimental design, results, and interpretations remain the sole responsibility of the authors.

References

Bedi, P. and Sharma, C. (2016). Community detection in social networks. *Wiley interdisciplinary reviews: Data mining and knowledge discovery*, 6(3):115–135.

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

Correa, D. C. and Levada, A. L. M. (2011). Finding Community Structure in Music Genres Networks. *Poster Session*.

Dolgikh, D. and Jelínek, I. (2015). Graph-based music recommendation approach using social network analysis and community detection method. In *Proceedings of the 16th International Conference on Computer Systems and Technologies*, pages 221–227, Dublin Ireland. ACM.

Fortunato, S. (2010). Community detection in graphs. *Physics Reports*, 486(3-5):75–174.

Grover, A. and Leskovec, J. (2016). node2vec: Scalable Feature Learning for Networks. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 855–864, San Francisco California USA. ACM.

Haupt, J. (2009). Last.fm: People-Powered Online Radio. *Music Reference Services Quarterly*, 12(1-2):23–24.

Kipf, T. N. and Welling, M. (2016). Variational Graph Auto-Encoders. arXiv:1611.07308 [stat.ML].

Porter, A., Bogdanov, D., Kaye, R., Tsukanov, R., and Serra, X. (2015). AcousticBrainz: A community platform for gathering music information obtained from audio. In *Proceedings of the 16th International Society for Music Information Retrieval Conference (ISMIR 2015)*, pages 786–792, Málaga, Spain.

Stutzbach, A. R. (2011). MusicBrainz (Review). *Notes*, 68(1):147–151.

Traag, V. A., Waltman, L., and Van Eck, N. J. (2019). From Louvain to Leiden: guaranteeing well-connected communities. *Scientific Reports*, 9(1).

Vlegels, J. and Lievens, J. (2017). Music classification, genres, and taste patterns: A ground-up network analysis on the clustering of artist preferences. *Poetics*, 60:76–89.

Zheng, M., Domanskyi, S., Piermarocchi, C., and Mias, G. I. (2021). Visibility graph based temporal community detection with applications in biological time series. *Scientific reports*, 11(1):5623.