Music Taste Shapes Friends but Not Communities in Last.fm

A Network Science Analysis of the Last.fm Social Network

Jonas Bjaerke
COMP0123
University College London

December 29, 2025

Abstract

Online social networks are often assumed to form through homophily, whereby individuals connect with others who share similar interests or backgrounds. In the context of music-based platforms such as Last.fm, this raises the question of whether shared musical preferences shape both local friendships and the global organisation of the network. In this study, we analyse the Last.fm social network using network science methods to examine how music taste and geographic location relate to network structure at different scales. We find that music taste similarity is elevated between directly connected users but decays rapidly with network distance, approaching a random baseline at distances comparable to the average shortest path length of the network.

At the community level, detected communities exhibit weak additional alignment in music taste beyond what is expected from local proximity, with the exception of a small number of tightly connected, musically specialised subgraphs. In contrast, geographic location is strongly associated with large-scale community structure, with several major communities dominated by users from a single country.

Taken together, these results suggest that music taste primarily shapes local connectivity rather than global community structure, while geographic factors play a more prominent role in organising the network at larger scales.

Contents

1 Introduction

1.1 Network Under Study

This project studies the social network formed by users of the online music platform Last.fm. The dataset is publicly available through the Stanford Network Analysis Project (SNAP) and represents a friendship network among Last.fm users [1]. In this network, nodes correspond to individual users and undirected edges represent mutual friendship relationships on the platform.

In addition to the social graph structure, the dataset provides node-level attributes, including music preference vectors which encode the full set of artists listened to by each user and allow direct comparison of music taste similarity between users. The data also includes geographic information in the form of a country label for each user.

1.2 Research Question

The central research question of this project is:

> **To what extent do music taste and geographic location shape local connectivity and community structure in the Last.fm social network?**

To address this overarching question, we explore the following sub-questions:

- Do detected communities using Louvain's algorithm correspond to similarity in music taste or geographic location?
- How does music taste similarity vary from direct neighbours to randomly selected user pairs?

Distinguishing between local homophily and community-level structure is important because the presence of similarity-based ties does not necessarily imply that networks organise into attribute-homogeneous groups. Online platforms such as Last.fm provide an ideal setting to test this distinction, as they combine explicit social links with rich behavioural data and relatively low costs of forming connections. This allows us to assess whether shared interests merely influence who becomes friends, or whether they shape the global architecture of the network.

2 Literature Survey

Several studies have examined homophily and community structure in online music-based social networks, particularly using data from Last.fm and similar platforms.

Early work on the Last.fm social network focused primarily on pairwise relationships and link prediction. Bischoff [2] analyses friendship links by comparing connected and non-connected user pairs using musical preference similarity, demographic attributes, and graph-based features. Their results show that friends tend to be more similar in music taste than random pairs, confirming the presence of music taste homophily at the edge level. However, they also report relatively low absolute similarity values, which they attribute to the sparsity of high-dimensional music preference vectors. Importantly, their study finds that graph-based features, such as common neighbours, are more informative for link prediction than taste-based features alone. While this work establishes the relevance of homophily for individual social ties, it does not investigate how these effects relate to global network structure.

Subsequent research has extended the analysis from individual links to community-level structure. Guidotti and Rossetti [3] study how personal music listening behaviour relates to detected communities in the Last.fm social network. Using community detection algorithms, they examine whether users within the same community exhibit similar music preferences. Their findings suggest that communities are not strongly homogeneous in terms of genre preferences, although users with broader and more diverse listening habits tend to connect with one another.

A similar analysis strategy is explored in an unpublished student coursework provided for this module, which analyses homophily and community structure in a Deezer friendship network using Louvain community detection and comparisons with randomised and explicitly homophilic network models [4]. By comparing genre distributions and cosine similarities across large communities, they conclude that music taste has little influence on community structure and that the observed network more closely resembles a randomised graph than a purely homophilic one. However, the study focuses primarily on large, coarse-grained communities and genre-level features, which may overlook homophily effects that operate at smaller network scales or at the level of local connectivity.

Taken together, existing research suggests that music taste homophily is present at the level of individual social connections, but that its influence on large-scale community structure is weak or limited. At the same time, several studies report stronger effects for non-taste attributes such as gender, age, and geographic location [2]. What remains less clear is how local homophily effects relate to community-level outcomes, and whether the observed weak community-level similarity reflects the aggregation of short-range effects across the network or something else.

3 Methodology

3.1 Data and Network Representation

We analyse the Last.fm social network dataset provided by the Stanford Network Analysis Project (SNAP) [1]. The network is modelled as an undirected graph $G = (V, E)$, where each node $v \in V$ represents a user and each edge $(u, v) \in E$ represents a mutual friendship relationship between users. The network contains $|V| = 7{,}624$ nodes and $|E| = 27{,}806$ edges.

Each node is associated with attribute data describing music taste and geographic location. Music taste is represented at the artist level. Let $\mathcal{A}$ denote the set of all artists appearing in the dataset, with $|\mathcal{A}| = 7842$. For each user u, we construct a binary feature vector $\mathbf{x}_u \in \{0, 1\}^{7842}$, where

$$(\mathbf{x}_u)_i = \begin{cases} 1, & \text{if user } u \text{ listens to artist } i, \\ 0, & \text{otherwise.} \end{cases}$$

This vector provides an artist-level representation of each user's music taste. To enable meaningful comparison across users with differing library sizes, we normalise each vector by its ℓ_1 norm so that the entries sum to one. Geographic information is provided as a country label c_u for each user, with 18 unique countries.

3.2 Basic Network Properties

We begin by computing standard descriptive network statistics to characterise the overall topology of the graph. These include the average shortest path length, the degree distribution, the average clustering coefficient, and degree assortativity. Together, these measures provide a baseline description of the network's topology.

3.3 Community Detection

To analyse the global structure of the network, we apply the Louvain community detection algorithm, which partitions the graph by maximising modularity. This procedure yields a set of communities $\{C_1, C_2, \ldots, C_k\}$ of varying sizes, capturing densely connected subgraphs relative to the overall network. These communities form the basis for subsequent analysis of both geographic composition and music taste similarity.

3.4 Community-Level Geographic Analysis

To evaluate whether detected communities align with geographic location, we analyse the country composition of each Louvain community. For each community C_j, we compute the proportion of users belonging to each country:

$$p_j(c) = \frac{|\{u \in C_j : c_u = c\}|}{|C_j|}.$$

This allows us to assess whether communities are geographically concentrated or instead consist of users from a wide range of countries.

3.5 Local and Community Music Taste Similarity

We examine homophily at a local scale by measuring music taste similarity between users as a function of network distance and comparing it to a random baseline. The aim is to assess whether neighbouring users exhibit greater similarity than would be expected from randomly connected pairs, and how this similarity changes as users become increasingly distant in terms of network hops within the social graph.

We sample user pairs and compute cosine similarity between music preference vectors for the following categories:

1. **Random pairs**: users sampled uniformly at random from the network.
2. **Distance-k pairs**: users separated by a shortest-path distance of k, for $k = 1, 2, \ldots, 9$.

At the community level, music taste similarity is measured both within and between Louvain communities. For each pair of communities (C_i, C_j), users are sampled from each community and cosine similarity is computed for sampled user pairs (u, v) with $u \in C_i$ and $v \in C_j$. The resulting averages provide estimates of within-community similarity $(i = j)$ and between-community similarity $(i \neq j)$.

Finally, we analyse the relationship between within-community similarity and the average shortest-path length within each community. This allows us to assess whether structurally compact communities exhibit noticable music taste similarity.

For each category, average similarity is computed over 500 sampled user pairs.

3.6 Tools and Reproducibility

All analyses were implemented in Python using NetworkX for network analysis and Matplotlib for visualisation. The full code and preprocessing steps are available at: https://github.com/jonasbjaerke/complex_network_analysis. Some procedures involve stochastic components for which random seeds were not fixed throughout; as a result, minor variations in numerical results may occur across runs, although all qualitative findings are robust.

4 Results

4.1 Network Properties

Table 1 summarises basic structural properties of the Last.fm social network. The network has an average degree of 7.294, and a low degree assortativity coefficient, suggesting little tendency for users to connect preferentially to others with similar degree. The average shortest path length of 5.232 is consistent with a small-world structure. The average clustering coefficient of 0.219 indicates a moderate tendency for users to form locally clustered groups.

Property	Value
Average degree	7.294
Degree assortativity coefficient	0.017
Average shortest path length	5.232
Average clustering coefficient	0.219

Table 1: Network properties of the Last.fm social network.

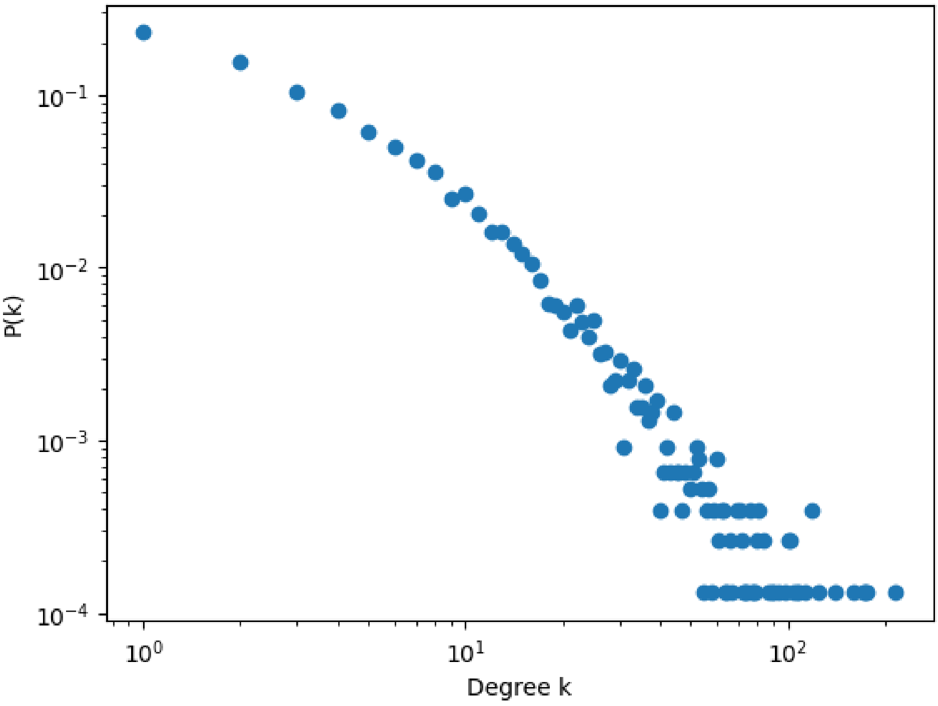

Figure 1: Degree distribution of the Last.fm social network (log scale) resembling power-law distribution.

4.2 Community Detection

Applying the Louvain community detection algorithm partitions the network into 27 distinct communities with a modularity of 0.8136, indicating a strongly modular structure. The resulting community size distribution is highly skewed, with a small number of large communities and many much smaller ones. To improve computational efficiency while retaining the majority of the network structure, subsequent analyses focus on the five largest communities, which together contain approximately 65% of all nodes in the network.

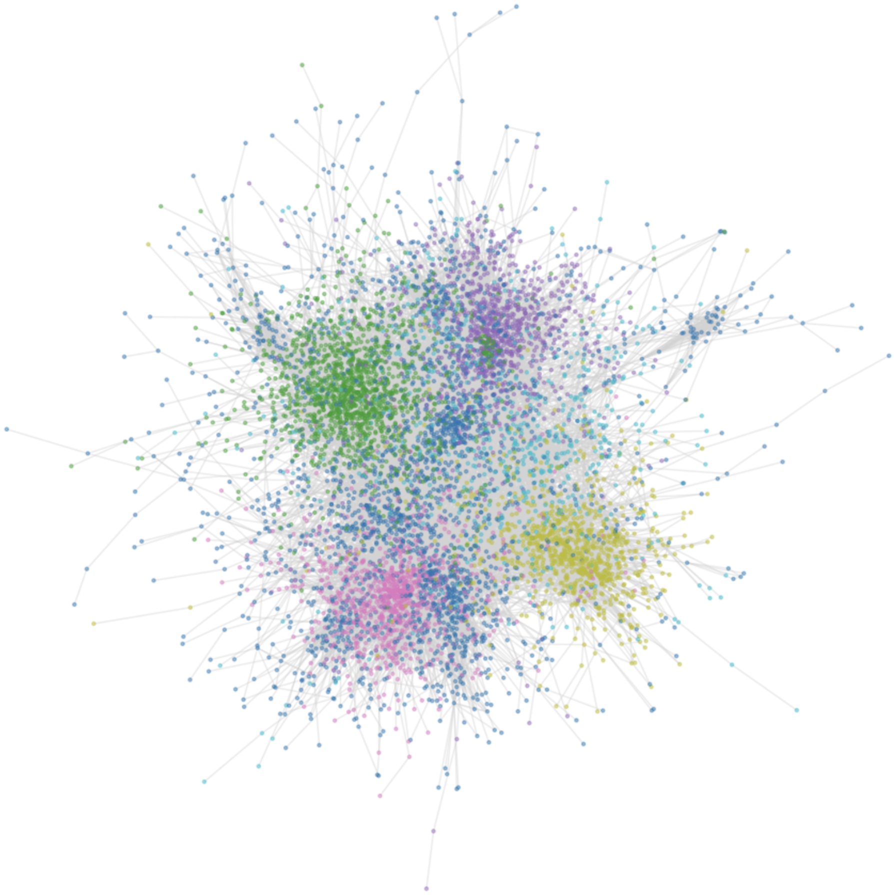

Figure 2: Visualization of the five largest Louvain communities in the Last.fm social network. Nodes belonging to the five largest communities are coloured, while all other nodes are shown in grey.

4.3 Community-Level Geographic Composition

Figure 3 shows the geographic composition of the five largest Louvain communities. Three of the five communities are almost entirely composed of users from a single country, with over 90% of nodes sharing the same country label, while the remaining two communities exhibit a more mixed geographic composition. This indicates that geographic location is closely associated with the formation of large communities in the network.

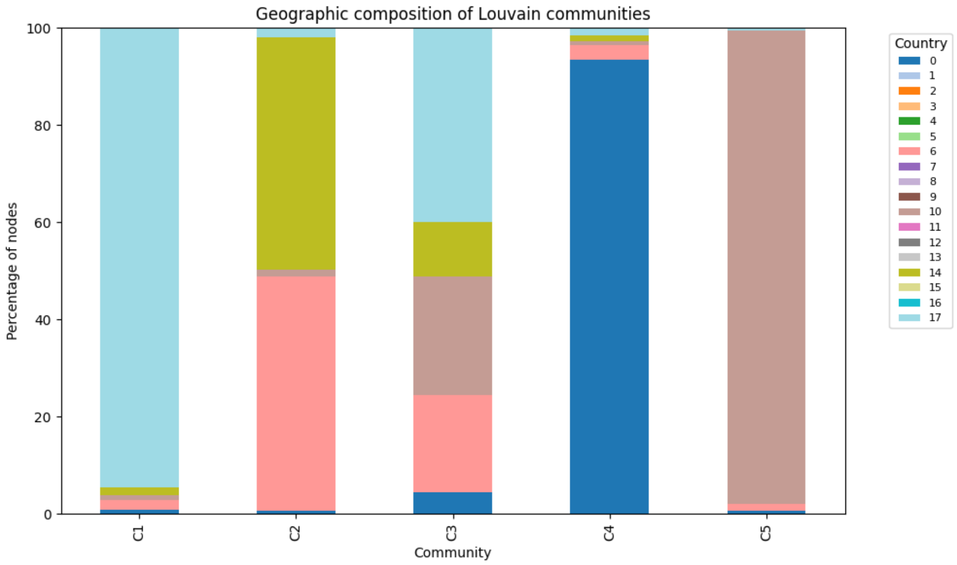

Figure 3: Geographic composition of the five largest Louvain communities.

To assess whether this geographic concentration could arise by chance, we construct a null model by randomly permuting country labels across nodes while keeping the network structure and community assignments fixed. Figure 4 shows the resulting community compositions under this randomisation. In contrast to the original network, no community is dominated by a single country, and the compositions closely resemble the overall country distribution.

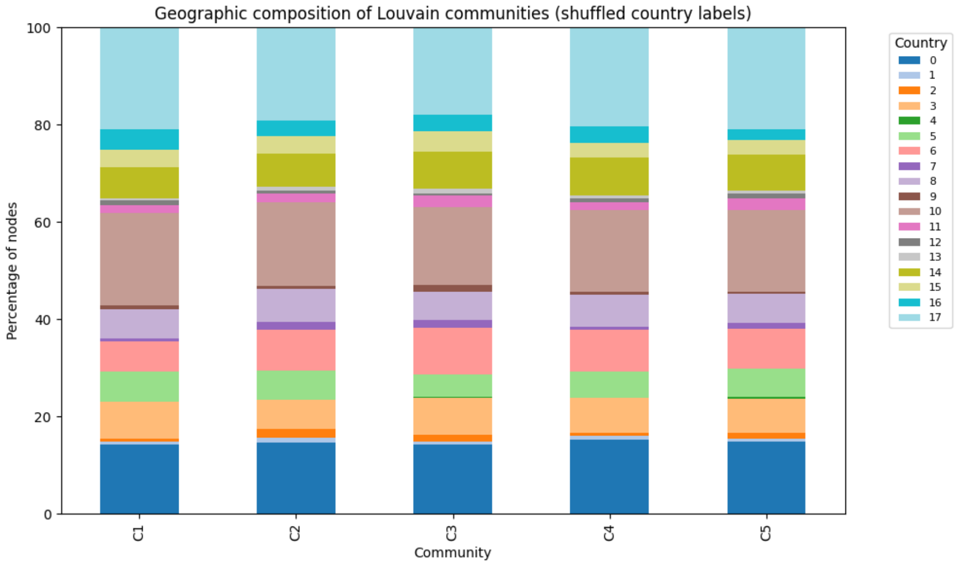

Figure 4: Geographic composition of the five largest Louvain communities under a null model in which country labels are randomly shuffled across nodes.

To further highlight the role of geographic location in community formation, we visualize the network using identical node positions but colouring nodes according to country of origin rather than community membership.

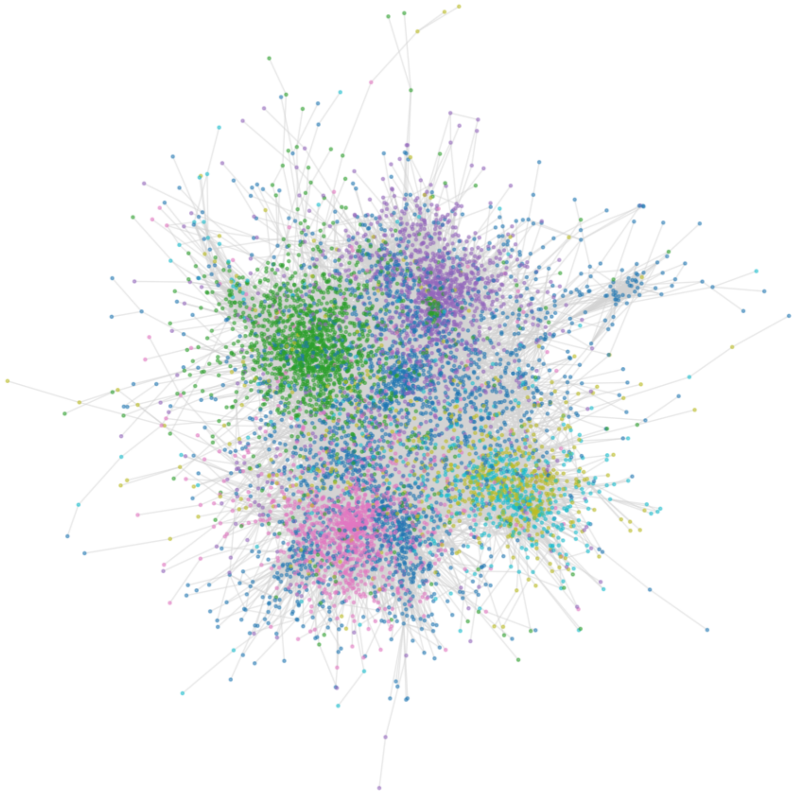

Figure 5: Visualization of the Last.fm social network with nodes coloured by country of origin. Nodes belonging to the five most common countries are coloured, while all other nodes are shown in grey.

Comparing Figures 2 and 5 shows a strong correspondence in the spatial organisation of the network. Large contiguous regions identified as single Louvain communities closely overlap with regions dominated by a single country, providing qualitative evidence that geographic location is closely alligned with community structure.

4.4 Local Music Homophily

Figure 6 shows how music taste similarity behaves on average as a function of graph distance between users. Users who are directly connected exhibit more than twice the average similarity of randomly selected pairs, indicating local music taste homophily. Notably, similarity decreases smoothly with increasing graph distance, approaching the random baseline by distance five, after which it appears to stabilise.

Interestingly, the average shortest path length of the network is approximately 5.2, which closely matches the distance at which music taste similarity becomes indistinguishable from random.

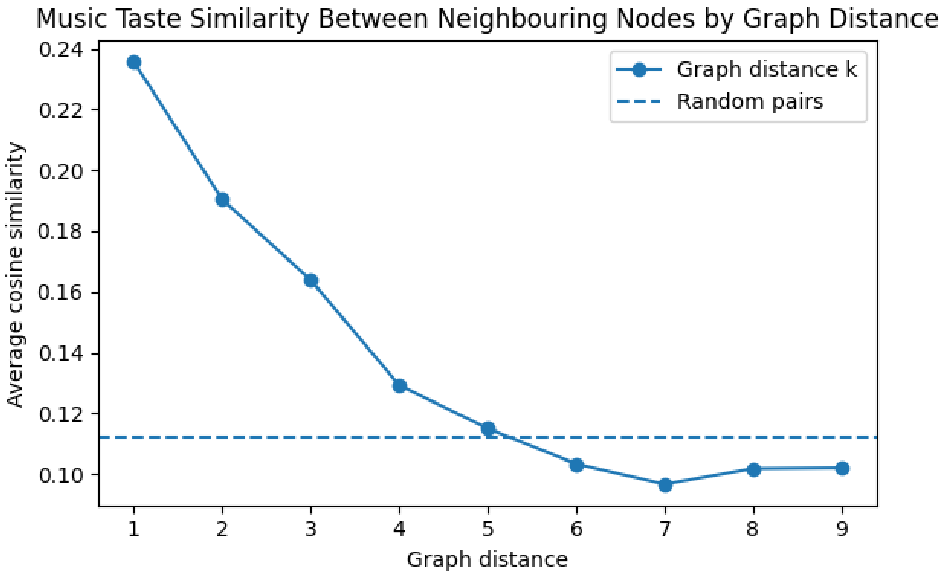

Figure 6: Average cosine similarity of music preference vectors for user pairs as a function of shortest-path distance. The dashed line indicates the average similarity for randomly selected pairs.

The result suggest that music preference plays a role at the local level, but that this effect does not persist beyond distances comparable to the average separation between users in the network. This observation motivates using the local similarity–distance relationship as a baseline for evaluating community-level patterns. The distance-based decay in similarity can be interpreted as the level of alignment expected to arise from local proximity and social interaction alone, for example through influence among friends. Comparing within-community similarity to this baseline therefore allows us to assess whether any communities exhibit levels of music taste similarity that are unusually high given their typical internal distances. Communities that lie well above the baseline cannot be explained by local proximity alone and may reflect the presence of additional organising mechanisms, such as tightly knit groups with shared external influences or common contexts that promote unusually strong similarity in music preferences.

4.5 Community-Level Music Similarity

Figure 7 shows the relationship between within-community music taste similarity and average shortest-path length across the 27 Louvain communities. Consistent with the local homophily results, a negative relationship is observed: communities with shorter internal path lengths tend to exhibit higher average music taste similarity.

However, this pattern does not fully align with the baseline implied by Figure 6. While most of the communities exhibit within-community similarity at levels comparable to those expected from their typical internal path lengths, a small subset lies well above this baseline, displaying substantially higher similarity than expected. These high-similarity communities stand out as very musically specialised, suggesting that music can play a decisive role in the formation of tightly connected groups

In contrast, communities with larger internal path lengths tend to be more musically heterogeneous, implying that their formation is less driven by shared music taste.

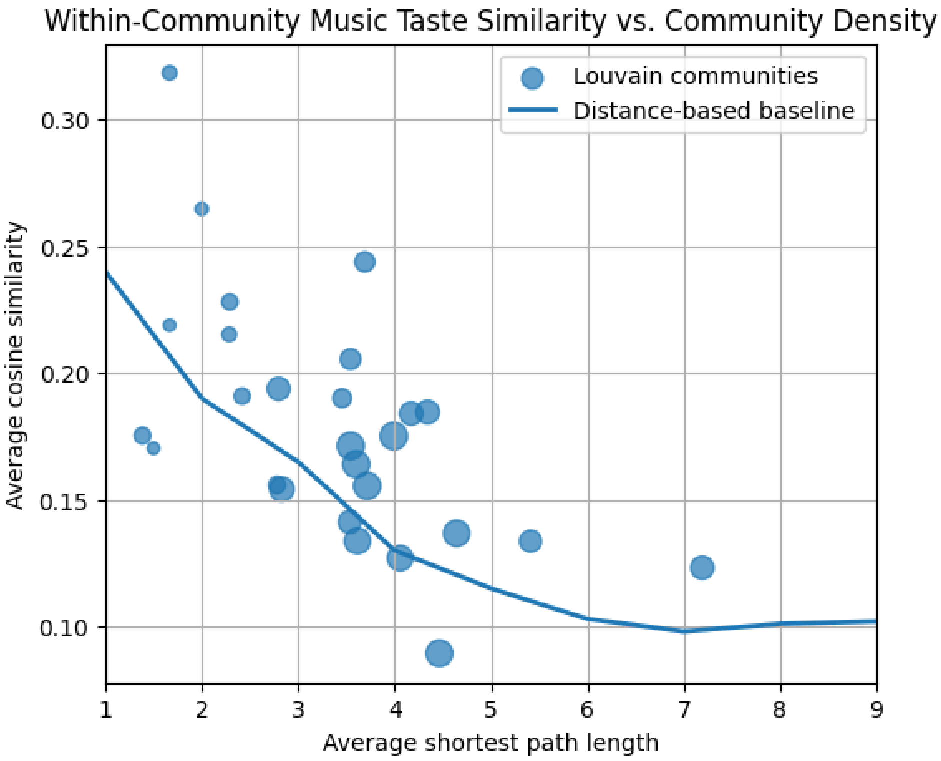

Figure 7: Each dot represents one of the 27 Louvain communities, with dot size proportional to the log of the number of users in the community. The baseline comes from figure 6

Table 2 reports music taste similarity across the five largest Louvain communities. The diagonal entries are slightly higher than the off-diagonal entries, indicating marginally greater similarity within communities than between them. However, the overall magnitude of these similarities is low and lies within the range implied by the baseline observed in Figure 6. This suggests that the elevated diagonal entries reflect baseline similarity arising from short path lengths within communities, rather than evidence of communities being organised around shared music preferences.

Consistent with this interpretation, the off-diagonal entries are close to the random-pair baseline, implying that music taste similarity between different large communities is largely indistinguishable from that of randomly selected user pairs.

	C_1	C_2	C_3	C_4	C_5
C_1	0.14				
C_2	0.12	0.16			
C_3	0.10	0.12	0.14		
C_4	0.12	0.15	0.11	0.17	
C_5	0.10	0.12	0.12	0.13	0.17

Table 2: Average cosine similarity between the five largest Louvain communities.

4.6 Supplementary Analysis: A Possible Feature Representation Bias

To assess potential issues arising from the binary feature representation of music taste, we examine how average cosine similarity varies with the number of liked artists for both connected and random user pairs (Figures 8). In both cases, we observe a positive linear trend, suggesting that users with larger music libraries are mechanically assigned higher similarity scores. Although it is somewhat trivial that users with broader music taste may naturally share more overlap with others, the binary feature representation does not account for preference strength and therefore one could argue this trend may reflect a representation bias, which should be taken into account.

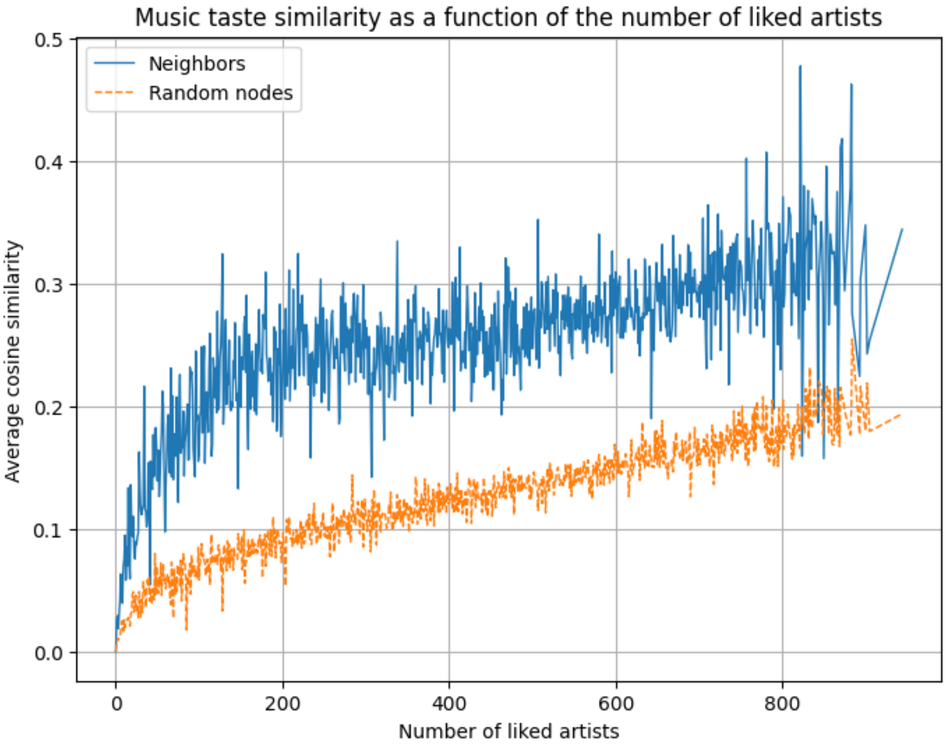

Figure 8: Average cosine similarity between users with a given number of liked artists (#positive entries in the binary artist vector) and their neighbours and random user pairs.

5 Discussion

5.1 Answer to the Research Question

This study examined the extent to which the topology of the Last.fm social network is correlated with music taste and geographic location. The findings indicate that both factors are related to network structure, but that they operate at different structural scales.

Music taste homophily appears to operate primarily at a local scale, consistent with prior studies [2, 5]. Elevated similarity is largely confined to immediate neighbours and does not persist across longer network distances, suggesting that music taste does not act as a global organising principle of the network. In practical terms, this suggests that users do not actively choose new friends based on shared musical preferences; instead, similarities in listening behaviour tend to develop through influence among friends who are already connected.

At the community level, average music taste similarity is only modestly higher within communities than between them, a pattern also reported in previous studies [4, 6]. Our analysis adds context by explicitly linking within-community similarity to average shortest path length. The observed similarity levels are largely consistent with a local similarity–distance baseline. This indicates that communities are not organised around shared music preferences, but rather that local similarity effects become weakly visible when averaged over denser network regions.

At the same time, a small subset of tightly connected communities exhibits within-community music taste similarity that is markedly higher than expected given their internal path lengths. These communities appear unusually musically specialised, suggesting that in some cases users do form communities around shared musical preferences. This is an interest finding which is not emphasised in prior studies of the Last.fm network, which largely report weak or absent taste-based community structure.

Lastly, geographic location shows a very strong association with large-scale network structure. Several of the largest communities are highly geographically concentrated, with some being represented solely by a single country. This suggests that location-based social ties continue to shape global network organisation, even on platforms designed around interest-based interaction.

5.2 Limitations

The supplementary analysis highlights a limitation of the chosen music taste representation. By relying on unweighted artist-level information, the similarity measure cannot distinguish between weak and strong preferences, treating all artists as equally important. This means that breadth of listening may be conflated with genuine similarity in musical taste. Access to weighted listening data, capturing how strongly users engage with individual artists, would allow for a more faithful representation of music taste and enable more robust similarity measures in future work.

References

[1] Stanford Network Analysis Project (SNAP). Last.fm social network. https://snap.stanford.edu/data/feather-lastfm-social.html, 2010. Accessed: 28 December 2025.

[2] Kerstin Bischoff. We love rock'n'roll: Analyzing and predicting friendship links in last.fm. 2012. Available at: https://dl.acm.org/doi/10.1145/2380718.2380725. Accessed: 28 December 2025.

[3] Riccardo Guidotti and Giulio Rossetti. "know thyself": How personal music tastes shape the last.fm online social network. 2019. Available at: https://pages.di.unipi.it/datamod/wp-content/uploads/sites/8/2019/10/DataMod_2019_paper_11.pdf. Accessed: 28 December 2025.

[4] RJHKI97. Exploring the relationship between community structure and homophily in music streaming social networks. Coursework report, COMP0123: Complex Networks and Web, 2023.

[5] Tomislav Duricic, Dominik Kowald, Markus Schedl, and Elisabeth Lex. My friends also prefer diverse music: Homophily and link prediction with user preferences for mainstream, novelty, and diversity in music. 2021. Available at https://domkowald.github.io/documents/2021asonam_lastfm.pdf, accessed 28 December 2025.

[6] Halil Bisgin, Nitin Agarwal, and Xiaoqing Xu. Does similarity breed connection? an investigation in blogcatalog and last.fm communities. 2010. Available at: https://www.researchgate.net/publication/220876142. Accessed: 28 December 2025.