Citation is not Collaboration: Music-Genre Dependence of Graph-Related Metrics in a Music Credits Network

Giulia Clerici* **Marco Tiraboschi***
Laboratory of Music Informatics (LIM)
Department of Computer Science
University of Milan, Italy
giulia.clerici@unimi.it marco.tiraboschi@unimi.it

ABSTRACT

We present a study of the relationship between music genres and graph-related metrics in a directed graph of music credits built using data from Spotify. Our objective is to examine crediting patterns and their dependence on music genre and artist popularity. To this end, we introduce a node-wise index of reciprocity, which could be a useful feature in recommendation systems. We argue that reciprocity allows distinguishing between the two types of connections: *citations* and *collaborations*. Previous works analyse only undirected graphs of credits, making the assumption that every credit implies a collaboration. However, this discards all information about reciprocity. To avoid this oversimplification, we define a directed graph. We show that, as previously found, the most central artists in the network are *classical* and *hip-hop* artists. Then, we analyse the reciprocity of artists to demonstrate that the high centrality of the two groups is the result of two different phenomena. Classical artists have low reciprocity and most of their connections are attributable to *citations*, while hip-hop artists have high reciprocity and most of their connections are true *collaborations*.

1. INTRODUCTION

A musician's path often crosses someone else's. As colleagues, two artists can jointly write a song, or one of them can feature in the other's track. On the other side, an admirer can cover a song of their idol, or a producer can remix their favourite hit. All these different types of relationships weave a network between artists. Musicians that either collaborate with many others or wrote frequently-covered songs behave as central hubs in this network.

Our goal is to determine the relationship between music genres and graph-related metrics in such a network. We want to determine if artists of different genres are more or less central and to highlight whether citations or collaborations are the dominant practices. In order to build this network, we analysed data from Spotify. This is a common choice in the literature because the platform is accessible to third-party applications via a REST Web API [1].

Some of these third-party applications can be useful tools for musicological analyses regarding music genres. Two examples are the web page *"Every Noise at Once"* [2] for the visualization of music genre similarities, and the web platform by Baratè and Ludovico [3] for investigating music genre labels. Also, at the moment of writing, it is the leading platform by the number of paying users [4].

Since Spotify provides an index of an artist's commercial success, many researchers investigated the correlation between this *popularity* index and other features, including audio [5], precomputed features [6], and also metadata, such as graph-related metrics [7].

The relationship between commercial success and graph-related metrics of music genres has been investigated by Oliveira *et al.* [8]. They built an undirected weighted graph in which nodes represent music genres. The weight of an edge between two genres is the number of hit songs on which two artists, each one associated with one of the two genres, have collaborated. They used Exploratory Factor Analysis to find latent variables correlated with the graph-related metrics and DBSCAN clustering to highlight different "collaboration profiles".

South *et al.* [9] investigated the behaviour of the eigenvector centrality for an undirected graph of artist collaborations. They observed a critical transition in the eigenvector centrality when removing the least popular artists from the graph. In the full graph *classical* music artists are the most central, but *hip-hop* artists emerge as the most central artists in the modified graphs. They explain this behaviour by introducing two types of social influence, which they formalize in a "Social Group Centrality" model.

Working on the same raw data as South *et al.*, we built a directed graph of artist credits. This is much truer to the nature of the data and allows for a finer analysis of the network structure. We found that their SGC model does not adapt as well to other types of centralities and analysed the graph connectivity to provide a structural explanation for the observed behaviour. We propose reciprocity as a much clearer discriminant feature between two different *crediting profiles*: *citations* and *collaborations*. This dichotomy explains in a more intuitive way the difference between *classical* and *hip-hop* artists, while it is completely lost in the undirected graph.

* All authors have contributed equally

2. THEORETICAL BACKGROUND

In this section we briefly detail theoretical concepts that are relevant to our research. We will use the following notation conventions. We define G as a directed graph, or *digraph*, determined by the set of its nodes V and the set of its arcs $E \subseteq V^2$. We let $N := |V|$ be the number of nodes in the graph. Without loss of generality, we consider the set V to be the set of integers between 1 and N. We call $A \in 2^{N \times N}$ the adjacency matrix of the graph, such that $A_{i,j} = \mathbb{1}\{(i,j) \in E\}$.

2.1 Reciprocity

The reciprocity of a digraph is a metric that quantifies how frequently, if there is an arc from node *i* to node *j*, there is also the arc from *j* to *i*. Garlaschelli and Loffredo [10] define reciprocity as the correlation coefficient between the entries in the adjacency matrix and the entries in its transpose, ignoring entries on the diagonal

$$\rho := \frac{\mathrm{Cov}\left[A_{i,j}, A_{j,i}\right]}{\mathrm{Var}\left[A_{i,j}\right]} \tag{1}$$

If the adjacency matrix is symmetrical, then $\rho = 1$. The digraph is perfectly reciprocal, and it could be represented as an undirected graph. If $A_{i,j} = 1 - A_{j,i}$ for $i \neq j$, then $\rho = -1$ and the digraph is unilaterally connected. If the covariance is 0, then $\rho = 0$ and arcs are reciprocated as often as they would if the same number of arcs were distributed at random in the graph.

2.2 Reachable Sets

In a digraph, the reachable set of a node is the set of nodes that are reachable from that node, i.e. nodes at a finite distance from it. The co-reachable set of a node is the set of nodes from which that node is reachable. The co-reachable set of a node in a digraph is the reachable set of that node in the transposed graph. The co-reachable set of node i is:

$$\mathcal{K}_i := \{j \in V \mid i \neq j \wedge d(j,i) < +\infty\} \tag{2}$$

where $d(j,i)$ is the distance from node j to node i.

2.3 Centrality Metrics

A centrality metric indicates the importance of a node in a network. The Spotify digraph is large enough that computing some centrality metrics is intractable. We are focusing mainly on geometric centralities, which can be approximated efficiently using HyperBall [11], and PageRank.

2.3.1 In-degree

One of the simplest and most intuitive measures for centrality is the *in-degree*, which is the number of incoming arcs of a node. The in-degree of node *i* is

$$c_i^{in} := \sum_{j=1}^{N} A_{j,i} \tag{3}$$

2.3.2 Closeness

Closeness centrality is based on the intuition that a node is more central the closer it is to all other nodes in the graph. The closeness of a node is defined as the reciprocal of the sum of the incoming distances from any other node.

$$c_i^{closeness} := \frac{1}{\sum_{j \in \mathcal{K}_i} d(j,i)} \tag{4}$$

The distances from non co-reachable nodes are ignored: their distance is infinite and the centrality would result to be zero. However, nodes with a small co-reachable set tend to have a high centrality value [12].

2.3.3 Lin Centrality

Lin [13] introduced a modified version of closeness centrality, that is weighted by the square of the cardinality of the co-reachable set.

$$c_i^{lin} := \frac{|\mathcal{K}_i|^2}{\sum_{j \in \mathcal{K}_i} d(j,i)} \tag{5}$$

2.3.4 Harmonic Centrality

Harmonic centrality [14] addresses the weaknesses of closeness, by taking the harmonic sum of the distances instead of the reciprocal of the sum.

$$c_i^{harmonic} := \sum_{j \in \mathcal{K}_i} \frac{1}{d(j,i)} \tag{6}$$

Harmonic centrality naturally ignores nodes outside the co-reachable set, because $\lim_{d \to \infty} 1/d = 0$.

2.3.5 PageRank

PageRank is a spectral measure of centrality. The vector of PageRank values for all nodes can be defined as the solution p to the following equation [12]

$$\begin{aligned} p &= \alpha p \bar{A} + (1-\alpha) v \\ p &\in [0,1]^N \mid \|p\|_1 = 1 \end{aligned} \tag{7}$$

The PageRank of a node can be interpreted as the probability distribution of ending a random walk on that node.

3. MUSIC CREDITS NETWORK

We analysed a graph of music credits obtained from Spotify data. To build the graph, we used the same raw data as South *et al.* [9]. It has been collected via the Spotify Web API [1] between December 2017 and January 2018 by exploring the network via breadth-first search.

The dataset contains 1 250 114 artists, which we represent as nodes in the graph. From now on, we will refer to nodes and artists interchangeably.

We built a digraph, where there is an arc going from node x to node y if there is a song in artist x's discography for which artists y is credited. We can read an arc going from x to y as "artist x credits y for one of their songs". In our digraph there are 7 435 330 arcs.

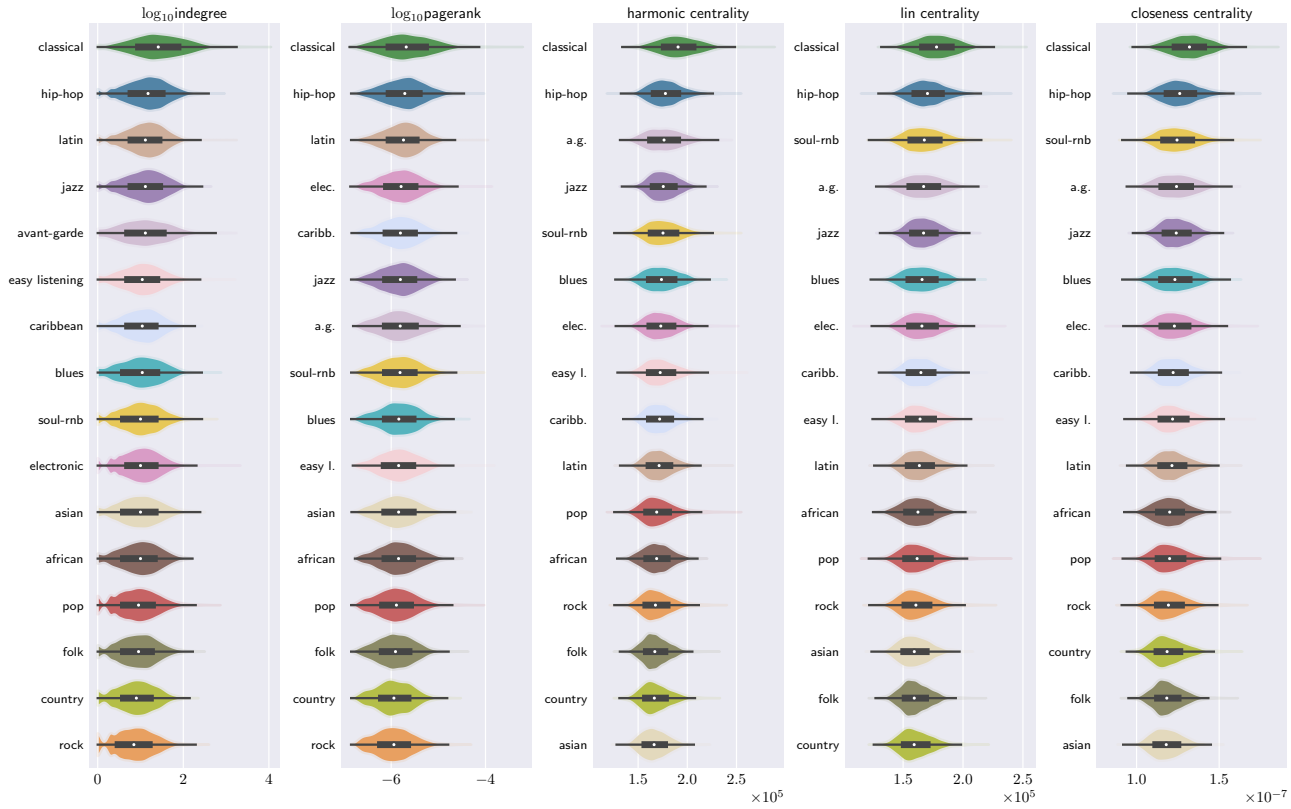

Figure 1. Violin-plot of the distribution of different centrality values conditioned on the musical genre. Genres are sorted in decreasing order of median value for each centrality. *Indegree* and *pagerank* are shown on a logarithmic scale.

3.1 Metadata

In the collected data there are nodes with no or partial metadata. We are mainly interested in two metadata: popularity and genre. Metadata was collected for 625 061 artists (around 50%). Popularity is a value between 0 and 100 related to the number, duration, and recency of streams of an artist's tracks [1].

One music genre or more can be associated to an artist. The number of artists with non-empty genre metadata in the dataset is 64 273 (around 10% of artists with metadata). The number of different genre labels is 1 533.

Following the definition of music genres as sets [15], we grouped them together in a total of 16 super-genres (i.e. super-sets), where an artist can belong to multiple genres. We know that this is an extreme oversimplification, but using all labels would result in more than a million pairs of genres to compare. Also, some sets have very few elements and this would affect negatively the statistical significance of the tests. Since there is no general consensus on the classification of genres in super-genres, we curated our own taxonomy. It was largely informed by AllMusic's genre classification [16], but we also consulted MusicMap [17], a "genealogy of popular music genres", and Every Noise at Once [2], a data-driven map of music genres.

3.2 Centrality Metrics

We analysed the distribution of centrality values for artists belonging to different music genres. We used WebGraph [18, 19], a Java library for the compression and analysis of very large graphs, to perform most of the computations. We used JPype [20] to interface WebGraph with Python, which we used for data visualization.

Figure 1 summarizes the results. We can observe that, for all the centrality metrics that we computed, the two music genres with the highest median values are *classical* and *hip-hop*, as previously found in the undirected graph. We assessed the significance of the differences between mean values using a Bayesian Student-T test [21], implemented in PyMC3 [22]. We defined the ROPE as an effect size between -0.1 and $+0.1$ (a "very small" effect size [23]) and set the significance threshold at $\alpha = 0.05$.

We can conclude that, for all centrality metrics, the mean value for *classical* artists is significantly greater than the mean value for all other genres. Also, the mean value for *hip-hop* artists is second for closeness centrality and Lin centrality. For harmonic centrality, in-degree, and PageRank, the comparison with the genre in third place is inconclusive, but the mean value for *hip-hop* artists is still greater than the mean value for all other 13 genres.

4. CREDITING PROFILES

We investigated the differences in crediting patterns between the two most central genres: *classical* and *hip-hop*. South *et al.* [9] proposed a model for social influence on the undirected graph of "collaborations". Their model does not explain the distribution of centrality values in our digraph. We propose reciprocity as an index of the differences of behaviour between the two genres: credits for

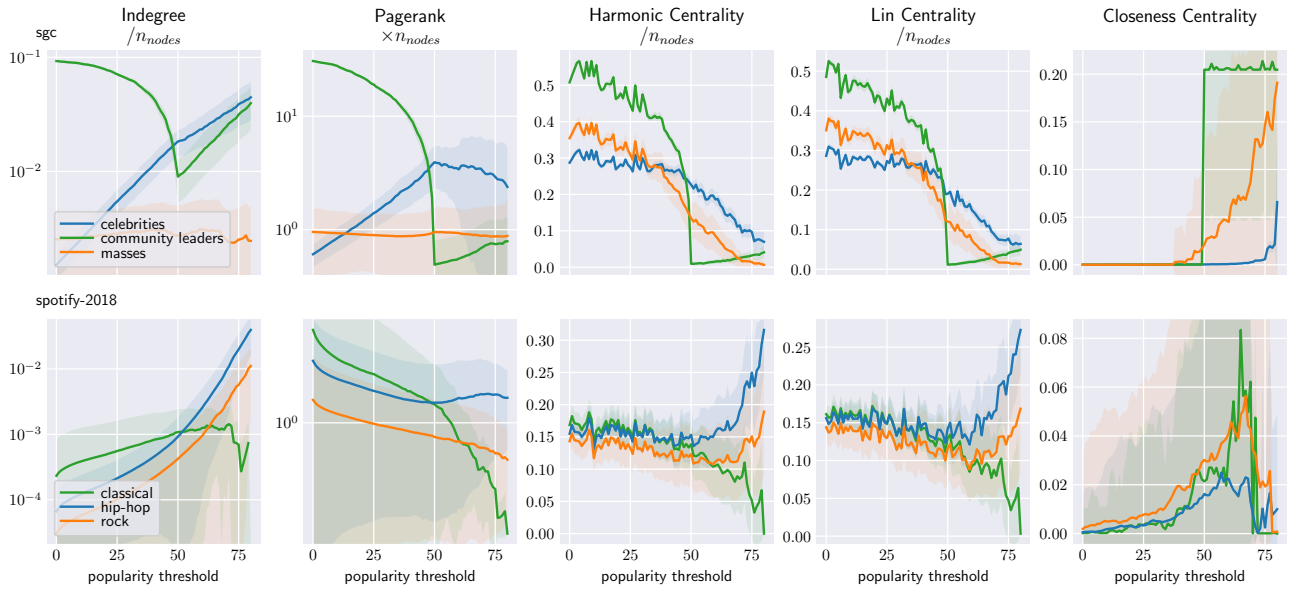

Figure 2. Transitions in node centralities under thresholding in the Spotify graph and in an SGC graph. On the x-axis are the popularity threshold values. Centrality values are normalized to remove trends naturally arising from changing the number of nodes in the graph (*in-degree*, *harmonic centrality*, and *Lin centrality* are divided by the number of nodes, *PageRank* is multiplied by the number of nodes). Solid lines are the average centrality values for nodes of one musical genre (or SGC class). Filled areas are between the average plus and minus 0.67 times the standard deviation (50% HDI for a normal distribution). *Indegree* and *pagerank* are shown on a logarithmic scale.

classical artists are more commonly *citations*, while for *hip-hop* artists they are more commonly *collaborations*. We want to emphasize the difference with the undirected graph in South *et al.* [9], where all edges are misinterpreted as collaborations.

4.1 Social Group Centrality

South *et al.* [9] observed that central *classical* artists have a large number of connections, especially with low-popularity nodes. On the other hand, central *hip-hop* artists are mostly connected to other artists with high popularity. They formalised this behaviour in a *Social Group Centrality* (SGC) model with three classes of nodes: *community leaders*, *celebrities*, and *masses*. *Community leaders* and *celebrities* are two cliques. Given a popularity threshold, *community leaders* are attached randomly to the *masses* nodes that have a popularity value below the threshold. On the contrary, *celebrities* are attached randomly to the *masses* nodes that have a popularity value above the threshold.

They verify their hypothesis introducing the concept of *thresholded graphs*: given the original graph with popularity labels on nodes and a threshold value, the popularity-thresholded graph is the sub-graph that only consists of nodes with popularity greater than the threshold. Nodes with no popularity metadata are not included in any thresholded graph. They analysed the average centrality of each genre for different threshold values and found that the eigenvector centrality is subject to a "critical transition". For the original graph, and for low thresholds, *classical* artists are the most central. After the popularity threshold value of 47 *hip-hop* artists become more central.

We implemented the SGC model in NetworkX [24] and compared the centrality *transitions* in a graph sampled from the model with our Spotify digraph. Figure 2 displays the average centrality values of three super-genres in the Spotify digraph and in the three classes of the SGC model. We can observe that, for most centralities, the SGC model graph and the Spotify digraph have different behaviours.

4.2 Reciprocity

We argue that the biggest distinction between "community leaders" and "celebrities" in the Spotify graph is that the former are highly cited, while the latter are truly collaborative. We propose node reciprocity to distinguish between *citations* and *collaborations*.

Reciprocity (as defined in Section 2.1) is a metric of the entire graph, but we are interested in investigating how much each node reciprocates arcs. Cheng *et al.* [25] define reciprocity over each pair of arcs: an arc is reciprocated if the arc between the same nodes, but in the opposite direction, exists. In our data, the direction of an arc is not entirely reliable: e.g. the out-degree of Mozart is 57, but that does not mean that Mozart featured 57 other artists' works in his own. In fact, those 57 nodes are mainly orchestras that performed Mozart's music. Wardil and Hauert [26] propose two indices to quantify a node's reciprocity: altruism (the balance between incoming and outgoing arcs) and activity (the number of arcs in either direction, normalized). However, having two indices is not practical.

We propose a node-wise reciprocity index, defined as Pearson's correlation coefficient between the entries in the adjacency matrix corresponding to incoming and outgoing arcs, which are the entries on the node's column and row, respectively. This takes after the graph reciprocity index

Genre	Q1	MED	Q3	IQR
african	0.739	0.889	1.000	0.261
asian	0.655	0.866	1.000	0.345
rock	0.655	0.866	1.000	0.345
avant-garde	0.615	0.866	0.970	0.355
folk	0.577	0.845	1.000	0.423
hip-hop	0.674	0.840	0.941	0.267
pop	0.623	0.833	1.000	0.377
latin	0.598	0.816	0.943	0.345
caribbean	0.577	0.816	0.935	0.358
jazz	0.539	0.804	0.939	0.400
soul-rnb	0.552	0.791	0.935	0.384
country	0.488	0.783	0.957	0.469
electronic	0.577	0.775	0.926	0.348
easy listening	0.479	0.747	0.926	0.447
classical	0.500	0.693	0.866	0.366
blues	0.365	0.655	0.913	0.548
Overall	0.583	0.816	0.949	0.365

Table 1. Summary of reciprocity values by super-genre: first quartile, median, third quartile and interquartile range.

introduced by Garlaschelli and Loffredo [10].

$$\rho_i := \frac{\mathrm{Cov}\left[A_{i,j}, A_{j,i}\right]}{\sqrt{\mathrm{Var}\left[A_{i,j}\right]\mathrm{Var}\left[A_{j,i}\right]}} \tag{8}$$

Empirically, it can be computed as

$$\hat{\rho}_i = \frac{\overleftrightarrow{a}_i - \overrightarrow{a}_i \overleftarrow{a}_i}{\sqrt{(1-\overrightarrow{a}_i)\overrightarrow{a}_i(1-\overleftarrow{a}_i)\overleftarrow{a}_i}} \tag{9}$$

where $\overrightarrow{a}$ is the normalized out-degree, $\overleftarrow{a}$ is the normalized in-degree, and $\overleftrightarrow{a}_i$ is the normalized number of reciprocated arcs.

$$\overrightarrow{a}_i := \frac{1}{N}\sum_{j=1}^{N} A_{i,j} \tag{10}$$

$$\overleftarrow{a}_i := \frac{1}{N}\sum_{j=1}^{N} A_{j,i} \tag{11}$$

$$\overleftrightarrow{a}_i := \frac{1}{N}\sum_{j=1}^{N} A_{i,j} \cdot A_{j,i} \tag{12}$$

In our graph, reciprocity is undefined for nodes with no outgoing arcs: in that case, since no edge is reciprocated, we define the reciprocity to be zero. It would be undefined for nodes with no incoming arcs, too, but there is no such node in our digraph, because of the data collection policy.

We summarize the distribution of node reciprocity for artists of different music genres in Table 1. We can observe that both the top two genres by median reciprocity (*African* and *Asian*) are music genres that are defined by their geographical origin. It would be interesting to further investigate the possible causes of these high reciprocity values.

But our main observation is that the reciprocity distributions for artists of *classical* music and *blues* have lower median values than others. The difference between the median values of *classical* and *hip-hop* artists is significant and practically relevant. This confirms our hypothesis that there is a preference for *citations* in *classical* music, while *hip-hop* artists prefer *collaborations*.

To give a qualitative insight, we sorted all the nodes by reciprocity and considered the ones with the highest popularity to find some examples that might be familiar to many people. Amongst the least reciprocating nodes, we can find many well-known artists. The top 10 artists in this sorting are: *Lil Pump*, *Green Day*, *Jorge & Mateus*, *Wham!*, *Oasis*, *Muse*, *Pearl Jam*, *Bruce Springsteen*, *Journey*, and *The Beach Boys*. We think that this may be due to the high number of cover songs that other artists published. On the other side, the top 10 most reciprocating nodes are less popular, both quantitatively and qualitatively. We report some exceptions: *Julian Casablancas* in tenth position, and *Guns N' Roses* in first position. The neighbours of *Guns N' Roses* are eight: five members of the band, two orchestras that recorded some of their songs, and one musician affiliated with one of the orchestras. What is unusual, in their case, is the fact that no unreciprocated covers had been uploaded at the time of the data collection by artists reached by the breadth-first-search crawler. This would probably be different if we repeated the experiment with updated data, or if we had the entire Spotify database.

5. CONCLUSIONS

We have shown that in music artists' networks built with data from music fruition platforms, arcs defined by credits do not imply collaborations. In light of this, undirected network models of collaborations can result unreliable, as they could be biased by the presence of citations.

Modelling credits with a digraph, we have shown that it is possible to quantify how much each artist is inclined to collaborate. We proposed a node reciprocity index for this purpose. Following the analysis of South *et al.* [9], we proposed reciprocity as the main difference between the *crediting profiles* of "celebrities" and "community leaders".

The concept of *classical* artists as "community leaders" is shown to be just a distortion introduced by the assumption that citations are collaborations. *Classical* artists in the network are actually highly-cited (famous composers) or highly-citing (orchestras) nodes. On the other hand, *hip-hop* artists, who were labelled as elitist "celebrities", score much higher reciprocity values.

5.1 Future Work

This result could find applications in recommendation systems, as in the automatic compilation of playlists. When building playlists of artists who belong to a community of collaborating musicians, the reciprocity index could help in filtering out false collaborators.

Also, we observed that the top two genres by median reciprocity are *African* and *Asian*, both of which are music genres that are defined by their geographical origin. It would be interesting to find a musicological explanation of this phenomenon.

6. ACKNOWLEDGMENTS

The authors would like to thank Tobin South for sharing the data they used for their study [9].

The authors would also like to thank Prof. Paolo Boldi for their valuable comments and helpful suggestions.

7. REFERENCES

[1] Spotify, "Spotify web api," https://developer.spotify.com/documentation/web-api, 2021, accessed: 2022-05-06.

[2] G. McDonald, "Every noise at once," https://everynoise.com, 2013, accessed: 2021-10-05.

[3] A. Baratè and L. A. Ludovico, "A web platform to extract and investigate music genre labels in spotify," in *Proceedings of the 19th Sound and Music Computing Conference*, 2022.

[4] Wikipedia, "Comparison of music streaming services," https://en.wikipedia.org/wiki/Comparison_of_music_streaming_services, 2023, accessed: 2023-02-06.

[5] M. Sciandra and I. C. Spera, "A model-based approach to spotify data analysis: a beta glmm," *Journal of Applied Statistics*, vol. 49, no. 1, pp. 214–229, 2022.

[6] C. V. S. Araujo, "A model for predicting music popularity on spotify," Extended Abstracts for the Late-Breaking Demo Session of the 21st International Society for Music Information Retrieval Conference, Montreal, Canada, 2020.

[7] Y. Matsumoto, R. Harakawa, T. Ogawa, and M. Haseyama, "Context-aware network analysis of music streaming services for popularity estimation of artists," *IEEE Access*, vol. 8, pp. 48 673–48 685, 2020.

[8] G. P. Oliveira, M. O. Silva, D. B. Seufitelli, A. Lacerda, and M. M. Moro, "Detecting collaboration profiles in success-based music genre networks," in *Proceedings of the 21st International Society for Music Information Retrieval Conference*. Montreal, Canada: ISMIR, 2020, pp. 726–732.

[9] T. South, M. Roughan, and L. Mitchell, "Popularity and centrality in spotify networks: critical transitions in eigenvector centrality," *Journal of Complex Networks*, vol. 8, no. 6, pp. 1–18, Dec 2020.

[10] D. Garlaschelli and M. I. Loffredo, "Patterns of link reciprocity in directed networks," *Physical Review Letters*, vol. 93, p. 268701, Dec 2004. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevLett.93.268701

[11] P. Boldi and S. Vigna, "In-core computation of geometric centralities with hyperball: A hundred billion nodes and beyond," in *2013 IEEE 13th International Conference on Data Mining Workshops*. New York, NY: IEEE, 2013, pp. 621–628.

[12] ——, "Axioms for centrality," *Internet Mathematics*, vol. 10, no. 3-4, pp. 222–262, 2014.

[13] N. Lin, *Foundations of Social Research*. New York, NY: McGraw-Hill, 1976.

[14] Y. Rochat, "Closeness centrality extended to unconnected graphs: The harmonic centrality index," Institute of Applied Mathematics University of Lausanne, Tech. Rep., 2009.

[15] F. Fabbri, "A theory of musical genres: two applications," *Popular music: critical concepts in media and cultural studies*, vol. 3, pp. 7–35, 2004.

[16] AllMusic, "Music genres," https://www.allmusic.com/genres, 2021, accessed: 2021-05-06.

[17] K. Crauwels, "Musicmap - the genealogy and history of popular music genres from origin till present (1870-2016)," https://musicmap.info, 2016, accessed: 2021-10-05.

[18] P. Boldi and S. Vigna, "The webgraph framework i: compression techniques," in *Proceedings of the 13th international conference on World Wide Web*. New York, NY: ACM, 2004, pp. 595–602.

[19] ——, "The webgraph framework ii: Codes for the world-wide web," in *Data Compression Conference, 2004. Proceedings. DCC 2004*, IEEE. New York, NY: ACM, 2004, p. 528.

[20] K. E. Nelson, M. K. Scherer, and U. N. N. S. Administration, "Jpype," 6 2020.

[21] J. K. Kruschke, "Bayesian estimation supersedes the t test." *Journal of Experimental Psychology: General*, vol. 142, no. 2, p. 573, 2013.

[22] J. Salvatier, T. V. Wiecki, and C. Fonnesbeck, "Probabilistic programming in python using pymc3," *PeerJ Computer Science*, vol. 2, p. e55, 2016.

[23] J. Cohen, *Statistical power analysis for the behavioral sciences*. Abingdon-on-Thames, UK: Routledge, 1988.

[24] A. A. Hagberg, D. A. Schult, and P. J. Swart, "Exploring network structure, dynamics, and function using networkx," in *Proceedings of the 7th Python in Science Conference*, G. Varoquaux, T. Vaught, and J. Millman, Eds. Pasadena, CA USA: SciPy2008, 2008.

[25] J. Cheng, D. M. Romero, B. Meeder, and J. Kleinberg, "Predicting reciprocity in social networks," in *2011 IEEE Third International Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third International Conference on Social Computing*, 2011.

[26] L. Wardil and C. Hauert, "Origin and structure of dynamic cooperative networks," *Scientific Reports*, vol. 4, no. 1, p. 5725, Jul 2014.