MUSICAL INFLUENCE NETWORK ANALYSIS AND RANK OF SAMPLE-BASED MUSIC

Nicholas J. Bryan and Ge Wang
Center for Computer Research in Music and Acoustics,
Department of Music, Stanford University
{njb, ge}@ccrma.stanford.edu

ABSTRACT

Computational analysis of musical influence networks and rank of sample-based music is presented with a unique outside examination of the WhoSampled.com dataset. The exemplary dataset maintains a large collection of artist-to-artist relationships of sample-based music, specifying the origins of borrowed or sampled material on a song-by-song basis. Directed song, artist, and musical genre networks are created from the data, allowing the application of social network metrics to quantify various trends and characteristics. In addition, a method of influence rank is proposed, unifying song-level networks to higher-level artist and genre networks via a collapse-and-sum approach. Such metrics are used to help interpret and describe interesting patterns of musical influence in sample-based music suitable for musicological analysis. Empirical results and visualizations are also presented, suggesting that sampled-based influence networks follow a power-law degree distribution; heavy influence of funk, soul, and disco music on modern hip-hop, R&B, and electronic music; and other musicological results.

1. INTRODUCTION

Network analysis has become a significant tool for understanding the dynamics of complex systems. Social network analysis, in particular, has increasingly garnered the attention of researchers across sociology, computer science, and statistics. Within the music information retrieval community, this has led to the creation of artist collaboration, recommendation, similarity, and influence networks.

Early music-based networks are found in Cano and Koppenberger [1] and Cano et al. [2]. Similarity networks from various online data sources are constructed with results showing the potential of how network analysis can help design recommendation systems. Further work of Jacobson [3, 4] and Fields [5,6] continued to show applications of automatic playlist generation, artist community detection, musicology, and sociology. Most recently, Collins investigated what is presumably the first computational analysis of musical influence using web scraping, web services, and audio similarity to construct influence graphs of a collection of synth pop music [7]. The work outlines the difficulty of constructing influence networks and motivates further investigation.

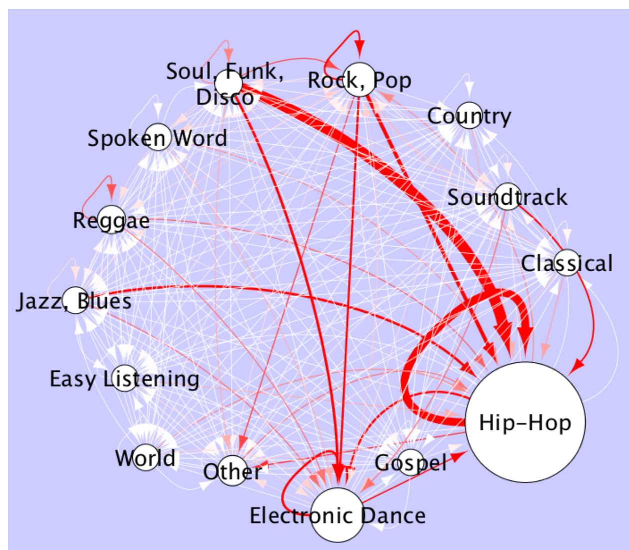

Figure 1. Visualization of Genre Flow. The size and opacity of a directed edge indicates the relative flow of samples from one genre to another.

The musicological and sociological impact of musical influence has considerable scope. Understanding how artists, musical styles, and music itself evolves over time can help us understand the creative process of music-making. Overall influence rank is also of considerable attraction, as music critics continually create top artist or producer lists within popular music (e.g. Rolling Stone Magazine). We present work towards this goal by studying influence found within sample-based music. [1] Directed influence graphs are constructed using a dataset from WhoSampled.com [8], a music website that chronicles sampling behavior via a community of contributors. Network analysis metrics and visualization such as Fig. 1 are employed on song, artist, and genre influence graphs in an effort to gain musicological understanding of the compositional act of sampling. In addition, a method of influence rank and analysis is proposed to help unify song-level networks to higher-level artist and genre networks via a collapse-and-sum approach. Empirical results found on constructed network graphs suggest musical influence-based networks follow a power-law degree distribution; heavy influence of funk, soul, and disco music on modern hip-hop, R&B, and electronic music; and various other anecdotal discussions of the unique corpus.

2. UNIQUE DATASET

The dataset was provided in agreement with WhoSampled.com and provides 42,447 user-generated records of sampling, excluding any entry involving cover song sampling. A baseline entry or *sample* of the dataset consists of a song-artist destination (who sampled the musical material) and song-artist source (source of the musical material sampled). In addition, other meta-data is provided, including destination and source release year, collaborating artists, featured artists, producers, genre, and part-sampled (i.e. vocals, drums, etc.).

For the purposes of this work, it is assumed that the large, high-quality dataset is a good representation of sampling behavior found within modern popular music and independent of any form of bias imposed by the user community. Labels of genre include hip-hop/R&B (H), electronic dance (E), rock/pop (P), soul/funk/disco (F), jazz/blues (J), reggae (R), country (C), world (W), soundtrack (S), classical (L), spoken word (K), easy listening (Y), gospel (G), and other (O). The part-sampled labels include: whole track (W), drum loop (D), bass line (B), vocals (V), hook (H), or other (O).

2.1 Genre & Part-Sampled Trends

To understand the data, we first take a look at the genre and part-sampled trends. The relative proportions of each genre are plotted in Fig. 2. Hip-hop/R&B, electronic dance, rock pop, and soul/funk/disco are dominate sources of musical samples, while hip-hop/R&B and electronic music are dominate destinations. The relative proportions and counts of each part-sampled are (W) 7.20% (3060), (D) 37.25% (15811), (B) 33.76% (14329), (V) 2.15% (913), (H) 17.25% (7321), (O) 2.39% (1013). Drum and bass components are the most dominant part-sampled followed by hook components.

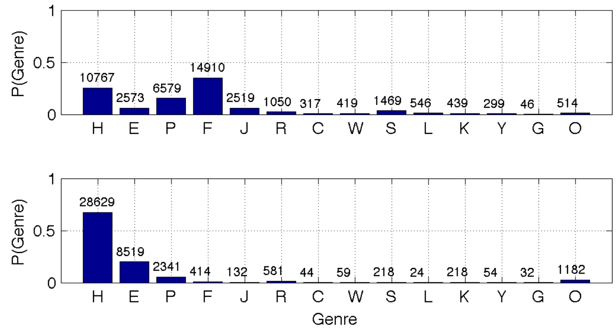

Figure 2. Source (upper) and Destination (lower) Genre Distributions with Absolute Counts.

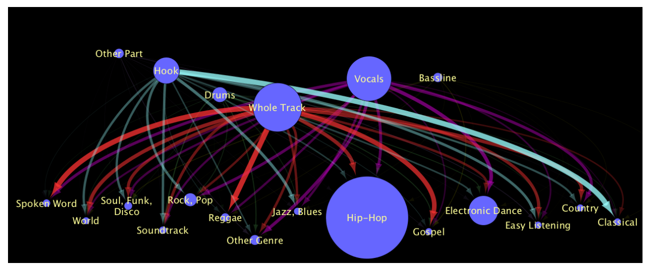

Figure 3. Visualization of Part-Sampled Flow. The size and opacity of a directed edge indicates the relative flow of part-sampled type to different genres.

Fig. 1 and Fig. 3 show more advanced visualizations emphasizing the flow of influence between genres [9]. Node size represents the destination proportions, while the directed edge opacity and thickness represent the conditional distribution of source genre given the destination genre. As seen, hip-hop/R&B consumes the most samples out of all genres, and within hip-hop/R&B most of the source material is from soul, funk, and disco as well as prior hip-hop/R&B material. In addition, it is also noticeable that electronic dance music more likely samples vocal material, while hip-hop/R&B more likely samples an entire portion of a song.

To measure how homogeneous the source material is for each destination genre, it is useful to employ the concept of *genre entropy H*, similar to discussions found in Jacobson [3] and Lambiotte [10]. Within this work, genre entropy is defined as

$$H_{g_k} = -\sum_{g_j \in \Gamma} P_{g_j|g_k} \log P_{g_j|g_k}, \qquad (1)$$

where g_k is the k^{th} genre in the set of genres Γ and $P_{g_j|g_k}$ is the probability of source genre g_j given the destination genre g_k. If genre g_k samples only from a single other genre g_j, the entropy will be zero. If destination genre g_k sam-

[1] Within this work, sample-based music is defined as a musical work that in borrows material from another musical source, whether it be a direct manipulation of a recorded sound or less direct transcribed material.

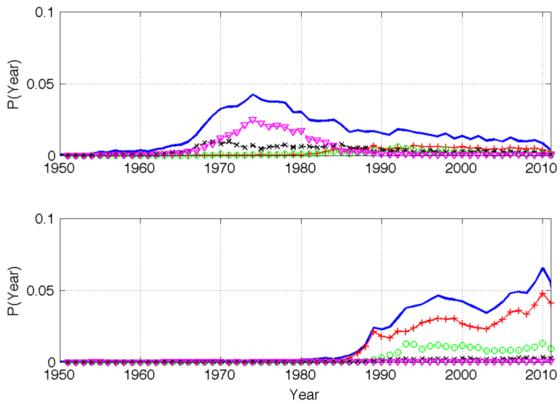

Figure 4. Distribution of Unique Samples Over Time. All samples (blue, solid), soul/funk/disco (magenta, triangles), hip-hop/R&B (red, plus), electronic dance (green, circle), rock pop (black, x) are shown across time for source material (upper) and destination material (lower).

ples uniformly from each source genre g_j, the entropy will be maximized. The genre entropy for the top five destination genres is shown in Table 1. The source samples used in

Genre	**Entropy** (bits)
electronic dance (E)	2.83
rock pop (P)	2.745
hip-hop/R&B (H)	2.356
soul/funk/disco (F)	2.242
reggae (R)	2.129

Table 1. Genre Entropy For Popular Destination Genres.

reggae music are the most homogeneous, while electronic dance music is the most heterogeneous. A closer look at electronic music reveals a near equal split of source material from hip-hop/R&B, electronic music, rock/pop, and soul/funk/disco with a slight preference towards the latter. Such evidence suggests differences in the creative process of sampling between genres.

2.2 Time-Based Trends

Initial observations of time-based trends are found when we view the proportion of samples per year within each genre. The trends can be viewed for both unique source and destination material normalized by the total instances of sampling as shown in Fig. 4. Plotting unique instances of source and destination material indicates general trends within each genre and eliminates the effect of a single popular sample swaying the proportions (as is the case without uniqueness enforced).

The general shape of the source material plot (upper) outlines the musical time frame of each genre (in terms of sampled source material), showing a rough outline of the rise and fall of soul/funk/disco and the rise of hip-hop/R&B. The general shape of the destination material (lower) outlines the increased popularity of sampling and/or listener trends within the WhoSampled.com user community. Interestingly, there is a sharp decrease in sample-based music centered around 2003. While further investigation is required, it interesting to note that this event directly coincides with the Recording Industry Association of America (RIAA) first litigation on Internet piracy and music copyright infringement [11]. Such legal policy would have created a more conservative and limited view of the musical practice of sampling, thus significantly affecting the music-making process.

3. NETWORK ANALYSIS

A discussion of network analysis, influence measures, and rank is presented with the motivation of observing how individual songs, artists, and genres influence one another. Complex network analysis provides significant tools for such characterization and begins with the formulation of a network graph. A graph $G = (N, E)$ is defined by a set of nodes N and edges E or equivalently an adjacency matrix $\mathbf{A}$. A weighted directed edge between node i and j is defined via $\mathbf{A}_{ij} = w_{ij}$ and 0 otherwise, where w_{ij} is the corresponding weight. For unweighted networks, all weights are either zero or one.

3.1 Degree Distributions

For a first general measure of how the music sample-based networks are constructed, degree centrality can be used to measure the influence from each node (song, artist, or genre) of a network. For a given node, the in- and out-degree centrality is defined as the respective in or out edge counts normalized by the total number of nodes $|N|$. The in- and out-degree distribution is then the proportion of degree $k = 1, 2, 3, ...$ nodes and can be used to characterize the network.

Power-law distributions $f(k) \propto k^{-\gamma}$ are an important family of distributions. Such distributions promote the concept of preferential attachment and are referred to as scale-free. To test the hypothesis that musical sampling follows a power-law, we can construct an unweighted acyclic song network using unique songs as nodes and sampling instances to create directed edges from destination to source. The in-degree distribution can then be computed and tested to follow a power-law distribution or not. Using methods described in [12], we find that the network is consistent with the hypothesis (p-value $= .16$ for $k \geq 3$ and $\gamma = 2.72$) and show the cumulative in-degree distribution in Fig. 5.

In terms of networks based on musical sampling, a scale-free network suggests the idea that very popular samples will only continue to increase in popularity. In addition, if

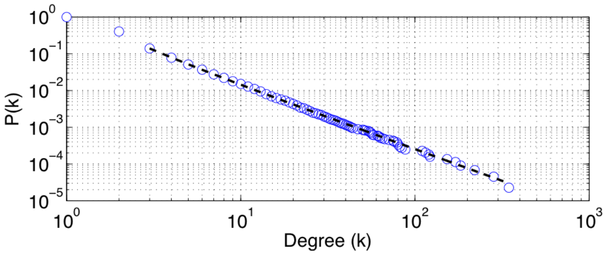

Figure 5. Cumulative In-degree Distribution $P(k)$ of the Sample-Based Song Network (log-log scale).

any of the very popular samples were to be removed, large portions of sample-based music would cease to exist (or at least be altered).

3.2 Influence Measures

To analyze and rank influence within each network, four closely related measures are commonly used: degree centrality, eigenvector centrality, Katz centrality, and PageRank. Degree centrality does not capture any indirect form of influence (as in the case of sample chains), motivating alternative methods. Eigenvector centrality extends degree centrality by weighting the importance of neighboring nodes to allow for indirect influence, but has limited application for acyclic networks [13]. [2] Katz centrality and PageRank appropriately modify eigenvector centrality. Both also provide a mechanism to capture indirect influence between nodes, compute an overall influence rank among each node, and observe the influence of one node to another. PageRank, however, down-weights influence created by a destination node that samples more than once, or in the case of artist nodes, down-weights influence from artists with lengthy careers. While this is desirable in numerous other contexts such as web search, we wish to equally weight each instance of sampling and restrict ourselves to Katz centrality.

The Katz influence matrix $\mathbf{I}_K$ is defined via

$$\mathbf{I}_K = (\mathbf{I} - \alpha\mathbf{A})^{-1} - \mathbf{I} \tag{2}$$

where $\mathbf{I}$ is an identity matrix, $\mathbf{A}$ is the adjacency matrix as before, and α is a decay factor which scales the indirect influence allowed to propagate though the network (larger α implies greater weight on indirect influence). This can be written in equivalent form as

$$\mathbf{I}_K = \alpha\mathbf{A} + \alpha^2\mathbf{A}^2 + ... + \alpha^k\mathbf{A}^k + ..., \tag{3}$$

where we can see that the influence is a weighted sum of the powers of the adjacency matrix [14]. When the values of $\mathbf{A}$ are zero or one, the powers of the adjacency matrix $\mathbf{A}^k$ have elements representing the number of sample chains of corresponding length k capturing various levels of indirect influence. For stability, $1/\alpha$ must be greater than the largest eigenvalue of A and for large networks, (2) becomes increasingly difficult to invert. Typically, only the overall influence rank is desired and is computed iteratively in a fashion to avoid a large memory footprint and matrix inverse required for $\mathbf{I}_K$.

For our purposes, it is desirable to have both the entire influence matrix and overall rank. Given $\mathbf{I}_K$, we can view the column of a node to find who influenced the node, or view the row of the node to find who the node influenced [15]. Summing the columns of the influence matrix produces a ranking of the most influential nodes, while summing the rows results in a ranking of the most influenced nodes. Such analysis is nicely suited for musicological analysis and motivates further improvements discussed below.

3.3 Collapse-and-Sum Influence Rank

For the given dataset, we would like to understand and analyze song, artist, and genre influence individually, as well as how each network relates to one another. To do so, individual networks can be constructed for song, artist, and genre networks with influence matrices and rank computed via (2) or (3). Building separate graphs, however, has several drawbacks. Most notably, there is no straightforward mechanism to relate the influence matrices of each network together appropriately. Furthermore, we would like to model the influence propagation on the song-level topology and then derive artist and genre influence measures, as the compositional act of sampling is presumably based on the musical material itself, rather than artist or genre connections.

To address this issue, a single influence matrix is constructed using the song-level network (see Section 3.1) and is used to create the artist and genre influence matrices, resulting in the proposed relational *collapse-and-sum* approach. To construct the artist-level influence matrix $\mathbf{I}_A$ from the song-level network, the song-level network is first used to compute the song influence matrix $\mathbf{I}_S$. Given $\mathbf{I}_S$, we then compute a derived artist influence matrix $\mathbf{I}_A$, knowing the source and destination song sets $S^s_{a_i}$ and $S^d_{a_i}$ belonging to each artist a_i. To do so, we take each artist a_i in the set of artists A and

- Sum over the destination song sets of each artist $S^d_{a_i}$, collapsing the appropriate columns of $\mathbf{I}_S$.
- Sum over the source song set of each artist $S^s_{a_i}$, collapsing the appropriate rows of $\mathbf{I}_S$.

The result of the process produces an artist-level influence matrix $\mathbf{I}_A$ which is directly derived from the song-level musical material, and is done so via linear combinations of the song-level influence matrix. The process can be duplicated

[2] The song network is exactly acyclic and the artist network is nearly acyclic.

or further reduced for other relations, such as artist-to-genre and song-to-genre influence.

Given the linear relationship from one network to the other, we can compute the relative proportions of influence between networks. As a result, we can analyze how influential a given song is to an overall artist's influence or how influential an artist is to a genre by taking ratios between the respective influence graphs, among other tasks. Secondly, by solely computing the influence on the acyclic, unweighted song-level network, we can compute $\mathbf{I}_S$ from a short, finite linear combination of the powers of the adjacency matrix without any iterative procedure and by knowing that the powers of the adjacency matrix $\mathbf{A}^k$, $k = 1, 2, 3...$ will go to zero when k is greater than the maximum sample chain length of the network. With sparse matrix representations, this modification can greatly reduced computation, increase the allowable in-memory network size, and additionally releases any restriction on α, allowing the user to choose any suitable weighting function. Application of this approach is found below in Section 4.

4. APPLICATION

Three levels of influence analysis and rank are computed for song, artist, and genre representations, providing a small-to-large inspection of the data. Various values of α are used to compare direct to indirect influence. For this purpose, (3) is rescaled to $\mathbf{I}_K = \mathbf{A} + \alpha^1\mathbf{A}^2 + ... + \alpha^{k-1}\mathbf{A}^k + ...$, allowing $\alpha = 0$ to only account for direct sampling, $\alpha = 1$ to equally account for direct and all indirect sampling, and values between zero and one to preferentially weight direct samples, but also account for indirect sampling.

4.1 Song Influence

The song-level influence matrix $\mathbf{I}_S$ is computed from the song network described in Section 3.1. The most influential songs are found in Table 2. We can observe the presence of many popular samples including "Change the Beat" by Fab 5 Freddy and the "Amen" break by The Winstons. It is particularly interesting to note that, for the "Amen" break, as α increases, the credit of influence intuitively moves from The Winstons to The Impressions, and finally to Jester Hairston. This is a result of a sample chain between material originating from Jester Hairston, that was first sampled by The Impressions, and then massively popularized by The Winstons.

4.2 Artist Influence

Starting with the song-level influence, we can collapse $\mathbf{I}_S$ to form an artist-based influence matrix $\mathbf{I}_A$. Table 3 shows the top influential artists.[3] We can also inspect the influence of an individual artist by looking at the corresponding row or column. Table 4, for example, names the top five influential and influenced artists of Jay-Z. Finally, we can also compute the relative proportion of influence created by each song within an artist's overall influence. The top three most influential songs of James Brown, for example, include "Funky Drummer" (14%), "Think (About It)" by Lyn Collins and produced by James Brown (9%), and "Funky President" (7.5%). Similar measures can be computed to indicate whether an artist gets more credit as a producer or performer.

James Brown (1.0)	James Brown (1.0)	James Brown (1.0)
Dr. Dre (0.34)	Dr. Dre (0.28)	Run-DMC (0.25)
Marley Marl (0.29)	George Clinton (0.25)	Fab 5 Freddy (0.23)[4]
George Clinton (0.28)	Marley Marl (0.25)	George Clinton (0.22)
Public Enemy (0.27)	Public Enemy (0.23)	Russell Simmons (0.19)
Rick Rubin (0.25)	Rick Rubin (0.22)	Kool & the Gang (0.19)
DJ Premier (0.25)	Fab 5 Freddy (0.22)	Marley Marl (0.18)
Material (0.24)	Material (0.21)	Rick Rubin (0.17)
Fab 5 Freddy (0.24)	Run-DMC (0.21)	Public Enemy (0.17)
Hank Shocklee (0.23)	DJ Premier (0.21)	Larry Smith (0.16)

Table 3. Artist Sample-Based Influence Rank for $\alpha = 0.0$ (left), $\alpha = 0.2$ (middle), and $\alpha = 1.0$ (right).

Influential ($\alpha = 0.2$)	**Influenced** ($\alpha = 0.2$)
The Notorious B.I.G. (0.97)	Girl Talk (1.0)
Dr. Dre (0.91)	Lil Wayne (0.80)
Puff Daddy (0.53)	The Game (0.53)
Nas (0.5)	DJ Premier (0.40)
James Brown (0.42)	Linkin Park (0.39)

Table 4. Top Influential and Influenced Artists of Jay-Z .

4.3 Genre Influence

The song-level influence matrix can further be reduced to a genre-based influence $\mathbf{I}_G$. The most influential genres found are: soul/funk/disco, hip-hop/R&B, rock/pop, jazz/blues, and electronic dance, while the top influenced genres are hip-hip/R&B, electronic dance, rock/pop, other, and reggae (for all values of α). Alternatively, the top songs and artist for each genre can also be computed (omitted due to space constraints).

5. CONCLUSIONS

An analysis of music influence and rank of sample-based music is presented using the WhoSampled.com dataset. General genre and time-based trends are found, identifying where and when the sampling source material is coming from as well as differences in how various genres are sampling others. Network graphs are employed to both understand general trends of sampling behavior, but to also find influence rank over songs, artists, and genre. A method of influence

[3] Entries with Fab 5 Freddy also include producers Material and Beeside. All three artists achieved high influence from "Change the Beat".

Change the Beat (Female Version) by Fab 5 Freddy (1.0)	Change the Beat (Female Version) by Fab 5 Freddy (1.0)	Change the Beat (Female Version) by Fab 5 Freddy (1.0)
Amen, Brother by The Winstons (0.82)	Amen, Brother by The Winstons (0.74)	Funky Drummer by James Brown (0.84)
Funky Drummer by James Brown (0.63)	Funky Drummer by James Brown (0.71)	Impeach the President by The Honey Drippers (0.62)
La Di Da Di by Doug E. Fresh (0.53)	La Di Da Di by Doug E. Fresh (0.51)	Synthetic Substitution by Melvin Bliss (0.55)
Think (About It) by Lyn Collins (0.49)	Impeach the President by The Honey Drippers (0.49)	Get Up, Get Into It, Get Involved by James Brown (0.54)
Impeach the President by The Honey Drippers (0.44)	Think (About It) by Lyn Collins (0.45)	The Big Beat by Billy Squier (0.51)
Funky President by James Brown (0.35)	Funky President by James Brown (0.37)	Scratchin' by The Magic Disco Machine (0.50)
Here We Go (Live at the Funhouse) by Run-DMC (0.34)	Synthetic Substitution by Melvin Bliss (0.36)	We're a Winner by The Impressions (0.46)
Bring the Noise by Public Enemy (0.33)	Here We Go (Live at the Funhouse) by Run-DMC (0.34)	Assembly Line by Commodores (0.46)
Synthetic Substitution by Melvin Bliss (0.32)	Bring the Noise by Public Enemy (0.32)	Amen by Jester Hairston (0.46)

Table 2. Song Sample-Based Influence Rank for $\alpha = 0.0$ (left), $\alpha = 0.2$ (middle), and $\alpha = 1.0$ (right).

rank is proposed, in an effort to unify higher-level artist and genre influence measures as appropriate linear combinations of song-level network influence. Empirical results suggest sample-based musical networks follow a power-law degree distribution; heavy influence of funk, soul, and disco music on modern hip-hop, R&B, and electronic music; and other musicological results.

6. ACKNOWLEDGEMENTS

This work was made possible from a generous collaboration with founder of WhoSampled.com, Nadav Poraz, who provided access to the rich and unique dataset. Additional help was provided by National Science Foundation Creative IT grant No. IIS-0855758.

7. REFERENCES

[1] P. Cano and M. Koppenberger: "The Emergence of Complex Network Patterns in Music Artist Networks," *In Proceedings of the International Symposium on Music Information Retrieval*, pp. 466–469, 2004.

[2] P. Cano, O. Celma, M. Koppenberger, and J. Martin-Buldú: "Topology of Music Recommendation Networks," *Chaos An Interdisciplinary Journal of Nonlinear Science*, 2006.

[3] K. Jacobson, M. Sandler, and B. Fields: "Using Audio Analysis and Network Structure to Identify Communities in On-line Social Networks of Artists," *In Proceedings of the International Symposium on Music Information Retrieval*, pp. 269–274, 2008.

[4] K. Jacobson and M. Sandler: "Musically Meaningful or Just Noise? An Analysis of On-line Artist Networks," *Computer Music Modeling and Retrieval. Genesis of Meaning in Sound and Music*, eds. S. Ystad, R Kronland-Martinet, and K. Jension. Springer-Verlag, Berlin, pp. 107–118, 2009.

[5] B. Fields, K. Jacobson, M. Casey, and M. Sandler: "Do You Sound Like Your Friends? Exploring Artist Similarity via Artist Social Network Relationships and Audio Signal Processing," *In Proceedings of the International Conference on Computer Music*, 2008.

[6] B. Fields, C. Rhodes, M. Casey, and K. Jacobson: "Social Playlists and Bottleneck Measurements: Exploiting Musician Social Graphs Using Content-Based Dissimilarity and Pairwise Maximum Flow Values," *In Proceedings of the International Symposium on Music Information Retrieval*, pp. 559–564, 2008.

[7] N. Collins: "Computational Analysis of Musical Influence: A Musicological Case Study Using MIR Tools," *In Proceedings of the International Symposium on Music Information Retrieval*, pp. 177–182, 2010.

[8] WhoSampled.com, Exploring and Discussing the DNA of Music, 2011. `www.whosampled.com`.

[9] P. Shannon, A. Markiel, O. Ozier, N. S. Baliga, J. T. Wang, D. Ramage, N. Amin, B. Schwikowski, and T. Ideker: "Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks." *Genome Research*, 13(11), pp. 2498-2504, 2003.

[10] R. Lambiotte and M. Ausloos: "On the Genre-fication of Music: A Percolation Approach (Long Version)." *The European Physical Journal B*, 50 (1-2), pp. 183-188, 2006.

[11] P. R. La Monica: "Music Industry Sues Swappers: RIAA Says 261 Cases Pursued for Illegal Distribution of Copyrighted Music; Amnesty Program Offered." *CNN Money, Technology News*, September 8, 2003.

[12] A. Clauset, C. R. Shalizi, and M. E. J. Newman: "Power-law Distributions in Empirical Data." SIAM Review 51(4), 661-703 (2009).

[13] M. E. J. Newman: *Networks: An Introduction*, Oxford University Press, New York, 2010.

[14] L. Katz: "A New Status Index Derived From Sociometric Analysis," *Psychometrika* 18, 1953.

[15] C. H. Hubbel: "An Input-Output Approach to Clique Identification," *Sociometry* 28 (4), pp. 377–399, 1965.