I SAID IT FIRST: TOPOLOGICAL ANALYSIS OF LYRICAL INFLUENCE NETWORKS

Jack Atherton and **Blair Kaneshiro**
Center for Computer Research in Music and Acoustics, Stanford University
{lja, blairbo}@ccrma.stanford.edu

ABSTRACT

We present an analysis of musical influence using intact lyrics of over 550,000 songs, extending existing research on lyrics through a novel approach using directed networks. We form networks of lyrical influence over time at the level of three-word phrases, weighted by tf-idf. An edge reduction analysis of strongly connected components suggests highly central artist, songwriter, and genre network topologies. Visualizations of the genre network based on multidimensional scaling confirm network centrality and provide insight into the most influential genres at the heart of the network. Next, we present metrics for influence and self-referential behavior, examining their interactions with network centrality and with the genre diversity of songwriters. Here, we uncover a negative correlation between songwriters' genre diversity and the robustness of their connections. By examining trends among the data for top genres, songwriters, and artists, we address questions related to clustering, influence, and isolation of nodes in the networks. We conclude by discussing promising future applications of lyrical influence networks in music information retrieval research. The networks constructed in this study are made publicly available for research purposes.

1. INTRODUCTION

Lyrics have been used to study many topics in music information retrieval (MIR) including genre classification [6], hit prediction [9], similarity searching [10], cultural studies [4], and computational musicology [5]. One approach to lyrical analysis is the *bag-of-words* model, which considers word frequencies in a text irrespective of order. In 2004, Logan et al. used this approach to produce promising preliminary results for measuring artist similarity through topic models, and observed that some genres naturally group with others based on shared vocabulary [9]. Fell and Sporleder later found that some genres (e.g. Rap, Metal) have relatively unique vocabularies, while others (e.g. Folk, Blues, and Country) cluster into groups [6]. Most recently, Ellis et al. computed bag-of-words novelty of lyrics and found that top-100 music was less lexically novel than less popular music [5].

In contrast, *n-gram* models consider ordered phrases of *n* words. A. Smith et al. used trigrams (n-grams with $n = 3$) and rhyme structures to develop a metric for cliché in lyrics, finding that number-one hits were more clichéd than average songs [14]; here, trigrams proved to be the better metric for measuring cliché. They also inspected their data by genre and found that genres had generally unique most-frequent phrases, though some phrases were shared by many genres. Later, Fell and Sporleder developed a suite of lexical features and used these, along with n-grams, to achieve performance gains in various classification tasks, but confirmed that n-grams alone achieved satisfactory baseline performance [6].

Networks—or graphs—are a natural and increasingly prevalent tool for analyzing structure between musical entities such as artists, songwriters, and genres. Networks comprise sets of nodes and sets of edges, or relations, between nodes. Most networks use unweighted, undirected edges, whereby edges are binary measures of whether two nodes are connected. Weighted edges ascribe varying importance to the relationships, and directed edges give each relationship a direction or flow. A 2006 study by R. Smith revealed the community structure of rappers by constructing a network between rappers who collaborated [15]. This study weighted edges by frequency of collaboration and found that different groupings such as large communities, music labels, and groups such as the famous Wu-Tang Clan emerged when varying percentages of the least significant edges were removed from the graph. We will refer to this process here as *edge reduction*. Collins later considered the flow of musical influence in synth-pop music [3], while Gunaratna et al. built a collaboration network for Brazilian musicians and composers [8]. Finally, Bryan and Wang constructed a directed graph connecting genres based on sampling of musical content, showing that Funk, Soul, and Disco heavily influence many modern popular genres [2].

Both lyrics and graphs allow us to ask deep questions about similarity, popularity, and interconnectedness in the music landscape. To our knowledge, no study has formed lyrics-based networks to analyze musical relationships formed directionally over time, although Fujihara et al. created a system for hyperlinking between lyrics and from lyrics to audio [7]. The current study combines lyrical analysis with graphs to observe influence of genres, artists, and writers, through networks formed by re-use of lyrics. Our networks' directional edges highlight lyrical influence over time, allowing us to address questions like that raised by A. Smith et al. of whether Pop music makes use of existing clichés or creates new ones [14]. We examine the topology of influence networks of genres, artists, and writers to quantifiably assess grouping behavior, robustness of connections, and centrality within the networks.

The remainder of this paper is structured as follows: first, we explain the formation of the influence networks and the computation of several of their properties. Next, we demonstrate and visualize the topology of the networks. Finally, we show how the basic building blocks of network properties can be used to address many outstanding musicological and MIR research questions.

2. METHODS

2.1 Data Sources

Past research has shown that n-grams (phrases) are superior to bag-of-words (vocabulary) for lyrical MIR tasks such as classification and computational musicology [6, 14]. With this in mind, we obtained intact lyrics data and artist/writer metadata via a signed research agreement with LyricFind, whose data were used previously in the Ellis et al. bag-of-words study on lexical novelty [5]. We additionally obtained primary genre and release date at the album level through the free iTunes Search API. [1] After filtering out songs with no lyrics, as well as those with no entry in the sparse iTunes dataset, we were left with 554,206 songs, collectively representing 42,802 artists, 95,349 writers, and 214 genres.

2.2 Constructing the Networks

The first step was to exhaustively measure the trigrams present in every song. Phrases were considered equivalent if they were cleaned to the same base phrase. We cleaned the lyrics using a procedure previously validated by Ellis et al. [5], avoiding stemming the words and using their rules for misspellings, alternate spellings, hyphens, and slang (modified to avoid expanding contractions and to correct a few inaccurate slang terms). We also filtered out *stopwords*—words too common to impart any lexical significance (e.g. pronouns and articles). We used a list of English stopwords from the Natural Language Toolkit, [2] augmented with many of the contractions ignored in the cleaning phase and misspellings of stopwords common to the LyricFind data. If a phrase was reduced in size due to stopword removal, we added an additional word and repeated the process until we obtained a cleaned trigram with no stopwords. This process allowed us to consider and match phrases originally longer than three words on the basis of only semantically significant words; for example, two four-word phrases that differed only by a specific pronoun (e.g. "he" / "she") would be matched after stopword removal. After initial results revealed spurious phrases created across lines of lyrics, we modified the algorithm to search for phrases only within lines of lyrics.

Next, songs were separated by year of release date in order to compute their phrases' term frequency-inverse document frequency (*tf-idf*). Tf-idf is a robust measure of significance common to information retrieval that increases an item's weight if it is common within its document and decreases its weight if it is common across the dataset [11]. Past MIR studies have used tf-idf for automatic mood classification of lyrics [16]—also in conjunction with rhyme information [17]—and to measure lexical novelty [5]. To capture the changing significance of phrases over time, we treated each individual year as a separate dataset. This way, the first person to use a phrase would have a significantly higher idf for that phrase than a person using it when it is already popular. Tf-idf is computed as in equation (1), where n_p is the number of occurrences of a phrase p in a song, n_s is the number of phrases in that song, s_y is the number of songs in a year y, and s_p is the number of songs in that year containing p.

$$\text{tf-idf}(p) = \tfrac{n_p}{n_s} \cdot \log(\tfrac{s_y}{s_p}) \tag{1}$$

We then constructed the three influence networks, one each for genres, artists, and writers. For every phrase in the dataset, we generated a list of all pairs of songs sharing that phrase. Pairs of songs released in the same year were ignored. This limit sets a minimum on the time difference necessary before a repeated phrase is considered influential, and also avoids forming links between potential duplicate entries in the dataset. For pairs of songs occurring in different years, we formed an edge from the earlier song to the later one. The edge's weight was the product of the tf-idfs of the phrase in both songs in order to capture the significance of the phrase in both years. Using song metadata, we then added the edge to the genre, artist, and writer graphs. For example, if a phrase was used in a Rock song in 1990 and in a Pop song in 1991, the resulting edge was drawn from the Rock node to the Pop node in the genre graph. If either song had multiple artists or writers, we added edges between all possible pairings. If multiple edges were added between two nodes, they were combined by summing their weights.

Next, *influence* scores were computed for each node of each graph. The influence I_i of node i is defined as the ratio of the sum of its outgoing edge weights (e_{ij}) to the sum of its incoming edge weights (e_{ji}):

$$I_i = \frac{\sum\limits_j e_{ij}}{\sum\limits_j e_{ji}} \tag{2}$$

Influence is thus a measure of the degree to which a node impacted future work or quoted previous work, but does not depend on the node's total volume of work.

Genre diversity scores were computed for each artist and writer as the number of genres they are credited in, divided by their total number of songs. This gives a measure

[1] http://apple.co/1qHOryr
[2] http://www.nltk.org/

of how many genres the person has contributed to without being skewed by their total number of contributions. [3]

Finally, each node's *self-reference* score was computed as the weight of the edge pointing from that node to itself, divided by that node's total number of contributions. This normalization again avoids skewing the score by volume, as edge weights were formed by summing over all possible pairings of phrases shared between two nodes.

The graphs were constructed with Graph-tool. [4] We make the graph data from this study publicly available for research purposes [1].

2.3 Network Analyses

The graphs were first assessed for clustering behavior. Adapting the method used by R. Smith, we analyzed the graphs in stages while removing increasing percentages of the least significant edges [15]. The first, *global* edge reduction method removed the $X_g\%$ lowest-weighted edges across the entire the graph, with X_g ranging from 0 to 99. The second, *local* method removed edges from each node that have a weight less than $X_l\%$ of the strongest weight at that node. At each stage of both procedures, the graphs were analyzed for their strongly connected components in order to examine the grouping behavior of the nodes.

Next, we performed *multidimensional scaling* (MDS) on the genre graph in order to embed its nodes in two dimensions for visualization. MDS converts a set of pairwise dissimilarities among objects into coordinates that can be used to visualize the objects in a low-dimensional space [13]. Here, the dissimilarities were computed using mutual influence D_{ij} below, where e_{ij} is the weight between nodes i and j. Graph visualizations were performed using Gephi. [5]

$$D_{ij} = \log(e_{ij} * e_{ji})^{-1} \tag{3}$$

Finally, we computed a series of correlations between the various metrics defined above, as well as the in-degree, out-degree, and average tf-idf of incoming and outgoing edges of each node. The r and p values were computed with the Scipy Statistics *pearsonr* function. [6]

3. RESULTS

3.1 Most Common Phrases

Table 1 shows the phrases used across the highest number of years. Many of these phrases are considered timeless, and all are semantically significant. Also, no phrase occurred in every year. The repetition of "dream(s) come true" does suggest that using word stems might improve performance. We note also that pronouns and other stopwords are absent from all phrases, which allowed the consideration of longer phrases with internal stopwords. For example, "makes feel like" is a combination of "makes (me) feel like," "makes (you) feel like," and other similar phrases. Overall, we treat these results as verification that our cleaning procedure was adequate.

Phrase	Years	Phrase	Years
dreams come true	49	one two three	43
never let go	48	late last night	42
new york city	46	whole wide world	42
long time ago	46	come back home	42
dream come true	45	makes feel like	41

Table 1: The most common phrases, ordered by number of years in which they appeared (maximum possible is 62).

3.2 Network Components

In our global edge reduction analysis, we expected that the graphs would split into several components. Instead, each graph remained concentrated in one large strongly connected component, with a few negligible side components. The size of the central component at a few points in the global edge reduction process is shown in Table 2.

The writer graph was the most robust: its central component was largest at nearly every point in the edge reduction process, and with only the top 1% of edges remaining in the graph ($X_g = 99$), it still contained nearly 200 components of 5 or more writers. We believe this result arose because many songs have multiple writers, while few songs have multiple artists; therefore, more relationships among writers would emerge from the same songs.

Using the local edge reduction method, a few small, significant components did break off of the main component. For example, with $X_l = 2$, the pairs {Celtic, Contemporary Celtic} and {Folk-Rock, Contemporary Folk} split off the main genre component. Table 3 shows the Brazilian and Spanish-speaking Latin American components formed with $X_l = 3$ and $X_l = 4$, respectively.

The graphs' strongly connected components split apart much more quickly using local edge reduction than with global edge reduction. At $X_l = 4$, the main component consisted of 21 genres; at $X_l = 20$, the only components remaining of size more than 1 were the two in Table 4. At $X_l = 25$, the main component had reduced to the pair {Rock, Pop}, but the Latin American component remained unchanged from $X_l = 20$. The answer to why the Latin American component was so robust probably lies in our data preparation method: since we did not filter out the stopwords of any language but English, the connections between genres of other languages were strengthened by spurious connections with no lexical significance. This result shows the importance of a cleaning procedure that works uniformly across the dataset.

Graph	$\mathbf{X_g = 0}$	$\mathbf{X_g = 90}$	$\mathbf{X_g = 99}$
Genre	95%	54%	13%
Artist	99%	52%	12%
Writer	98%	66%	28%

Table 2: Percentage of nodes in the central component with $X_g\%$ edge reduction.

[3] Un-normalized genre count did not significantly interact with any other variables in the study.

[4] https://graph-tool.skewed.de/

[5] https://gephi.org/

[6] http://docs.scipy.org/doc/scipy-0.16.0/reference/

$\mathbf{X_l = 3}$	$\mathbf{X_l = 4}$
MPB	Latin Pop
Sertanejo	Latin Alternative & Rock
Samba	Latino
Pagode	Salsa y Tropical
Axé	Regional Mexicano
	Baladas y Boleros
	Latin Urban

Table 3: Genres contained in two components that split from the main component with $X_l\%$ edge reduction.

Main Component	Latin American Component
Pop	Latin Pop
Rock	Latin Alternative & Rock
Alternative	Latino
R&B/Soul	Salsa y Tropical
Country	Regional Mexicano

Table 4: Components with $X_l = 20\%$ edge reduction.

3.3 Multidimensional Scaling and Visualizations

To explore the seemingly central nature of the graphs, we performed MDS and visualized the genre graph.[7] The visualization (Figure 1) promisingly showed that nearly all edges in the graph were focused toward the center.

To more closely observe the behavior at the center of the graph, we next visualized only the 28 genres that appear in the central component with global edge reduction ($X_g = 99$), and displayed only each node's top three incoming edges. Using dissimilarities computed from all edges (not just those present in the visualization), we performed MDS again to obtain the node positions for Figure 2 and Figure 3. Jazz, Pop, and Rock are firmly at the center of the genre network. Here, 21 of the 28 nodes in the central component include Jazz among their top three influences, while 16 include Pop and 13 include Rock.

3.4 Influence and Self Reference

We next sought a statistical explanation for the centrality of the highlighted nodes in Figure 1. Turning first to influence, the ratio of a node's outgoing to incoming edge weights, we expected that central genres would have high influence, meaning that many genres draw from the phrases used in the central genres. In fact, the extremes of the influence metric are dominated by outliers, including rare genres as well as artists and writers who appear only very early or very late in the dataset. These groups do not have much opportunity for incoming or outgoing edges, respectively. In contrast, central genres are referenced at about the same rate that they reference previous material, having influence scores close to 1.0. Figure 2 and Table 5 show the influence of central genres.

We then turned to the self-reference score, a measure of how much a genre re-uses lyrics from its past. Our intuition here was that genres that refer to themselves frequently create a particular subculture that is ripe for other genres to draw influence from. Table 5 shows the top 10

[7] Because of their sheer size, visualizations of the artist and writer graphs were not feasible at the time of the study.

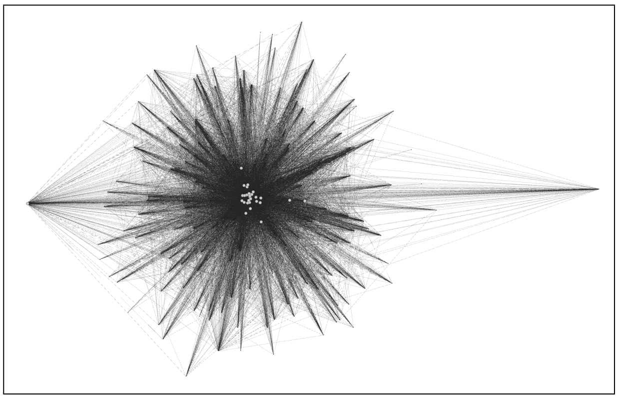

Figure 1: MDS layout of the genre graph, having 214 nodes and 17,438 edges. World, far left, is pushed out of the central area by its extreme dissimilarity to Pop Punk, far right. Light gray nodes are the 28 nodes in the central component with global edge reduction, $X_g = 99\%$.

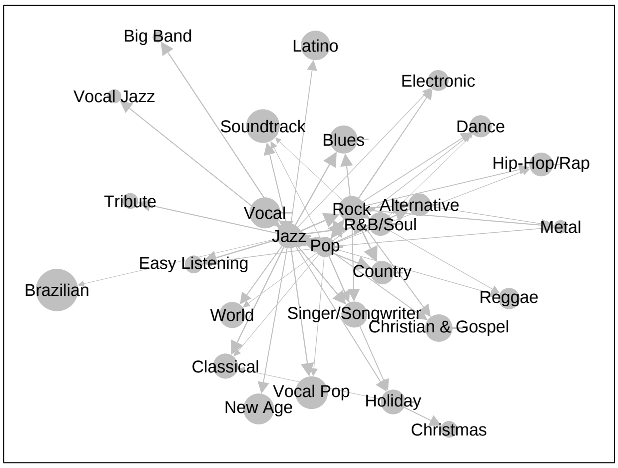

Figure 2: MDS layout of the central component of the genre graph. Node size denotes node influence, and arrow size denotes edge weight. Here, the sensitivity of the influence metric to outliers is shown (for example, with the large size of the Brazilian node).

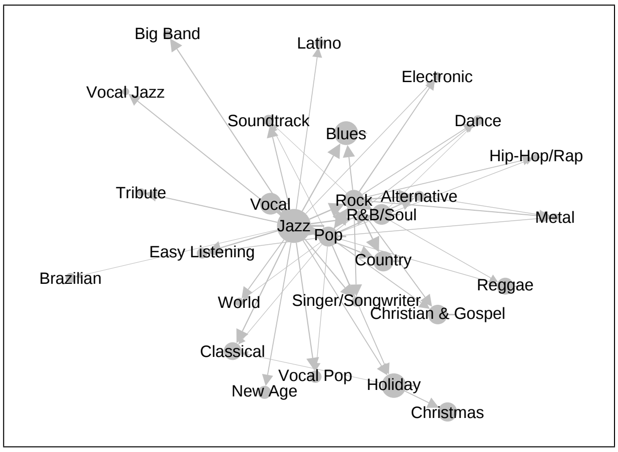

Figure 3: MDS layout of the central component of the genre graph. Node size denotes self-reference, and arrow size denotes edge weight. Self-reference, more than influence (Figure 2), correlates with centrality in the network.

Genre	Influence (%)	Self-reference (%)
Jazz	1.039 (65.3%)	**166.516 (100%)**
Holiday	1.031 (64.5%)	82.276 (99.6%)
Blues	1.176 (73.6%)	7.698 (99.1%)
Vocal	**1.301 (76.6%)**	6.818 (98.6%)
R&B/Soul	0.987 (63.2%)	6.369 (98.1%)
Country	0.974 (60.2%)	6.202 (97.7%)
Rock	1.083 (67.5%)	6.124 (97.2%)
Pop	0.836 (51.9%)	6.074 (96.7%)
Christian & Gospel	1.149 (71.9%)	5.883 (96.3%)
Christmas	**0.690 (47.6%)**	**5.759 (95.8%)**

Table 5: Top genres by self-reference. Raw values and percentiles are shown.

genres by self-referential behavior. These genres indeed correspond with genres near the center of the graph, and they are all in the central component of Figure 1. Jazz and Holiday have particularly outlying scores, perhaps reflecting that these genres often consist of *standards*—covers of widely known songs. Across the entire genre graph, log-self-reference correlates with centrality in the MDS graph (measured as the negative sum of euclidean distances from a node to all others), with $r = 0.610$, $p < 0.0001$.

After seeing the importance of self-reference in the genre graph, we computed this metric for the other two graphs. The list of top artists by self-reference is dominated by Jazz artists. This could reflect more the outlying nature of the genre than it does anything about Jazz artists. However, viewing the top artists by decade yields some interesting results when Jazz artists are ignored. For example, Sam Smith, Two Door Cinema Club, and Owl City are among the top-10 most self-referential artists of the 2010s.

3.5 Self-Reference and Volume

To better understand the metrics of self-reference and influence, we assessed their correlations with other aspects of the data. First, we determined the extent to which genre volume (number of songs in a genre) affects its self-reference. Using the un-normalized value of self-reference (i.e. the genre's raw self edge weight, not divided by the genre's volume), log-self-reference correlates highly with log-volume ($r = 0.846$, $p < 0.0001$). This is because the raw edge weight is a sum of all connections between songs of that genre, and more songs allow more connections. But, when self-reference is normalized by genre volume (see § 2.2), log-self-reference still correlates highly with log-volume ($r = 0.780$, $p < 0.0001$). Our intuition for this is that as the number of songs in a genre increases, the average quality of self-references increases, and so the normalized contribution from each song increases.

3.6 Genre Diversity, Influence, and Connectedness

Having investigated self-reference as a measure of phrase sharing within genres, we next assessed sharing of phrases across genres. We expected that people with high genre diversity are able to transfer phrases between genres, which would increase their influence score as the transferred phrases are referenced by people with less genre diversity.

Name	Influence (%)	Self-Reference (%)
Paul McCartney	0.970 (55.4%)	114.061 (99.8%)
John Lennon	0.974 (55.6%)	**119.660 (99.8%)**
Max Martin	**0.421 (31.7%)**	**0.739 (70.4%)**
Mariah Carey	1.039 (59.6%)	3.595 (88.9%)
Barry Gibb	**1.111 (62.4%)**	7.661 (95.0%)

Table 6: Top writers, ordered by number-one singles. Raw values and percentiles are shown.

However, we found that influence has a low, though statistically significant correlation with genre diversity ($r = 0.079$ for artists, $r = 0.155$ for writers, $p < 0.0001$).

Surprised by this result, we investigated further the effect of genre diversity on connections in the network. First, we investigated whether drawing influence from many genres correlates with more complex references and found a low, though statistically significant, correlation between writers' log-average incoming tf-idf value and genre diversity ($r = -0.167$, $p < 0.0001$). Next, we examined the degree to which writing in many genres correlates with directional influence forward and backward in time. We actually found a negative correlation between log-out-degree and genre diversity ($r = -0.572$, $p < 0.0001$), as well as between log-in-degree and genre diversity ($r = -0.563$, $p < 0.0001$). Thus, although influence (the ratio of outgoing to incoming edges) explains little genre diversity, increased genre diversity correlates moderately with less robust connections within the graph in both future and past directions. This could suggest that writers who contribute to a wider variety of genres use complicated phrases that are less likely to be shared with other writers, or that they use more stopwords that are filtered by the algorithm.

Artists showed similar correlation behavior to writers, but with lower correlation magnitudes, perhaps reflecting that artists are often a step removed from writing lyrics and may perform lyrics written by a variety of writers.

3.7 Top Genres, Writers, and Artists

We showed in § 3.5 that the volume of a genre in the dataset correlates highly with its self-reference score. Compared to other popular music genres, Rap has a particularly low self-reference score: it is the 6th most numerous genre in the data, but ranks 48th in self-reference. Similarly, Metal is the 8th most numerous genre in the data, but ranks 40th in self-reference. Rap's low self-reference score may reflect a particular subculture within this genre that values lyrical originality over references to past material.

Having analyzed lyrical influence between genres over time, we can now address whether Pop music is more clichéd than other genres because it draws from many sources or because it popularizes new phrases [14]. Rock and Pop are the two most common genres in our dataset. Rock's influence score is 1.083, while Pop's is only 0.836. Since Pop's influence is less than 1.0, Pop music quotes phrases from other genres more often than it influences them. This suggests that Pop music draws on existing clichés more than it creates new ones, especially when compared to other popular genres such as Rock.

Name	**Influence**	**Self-Reference**
The Beatles	**0.140 (21.5%)**	2.699 (95.4%)
Elvis Presley	0.847 (56.1%)	**9.964 (99.6%)**
Mariah Carey	1.688 (73.0%)	3.362 (96.6%)
Rihanna	0.381 (36.9%)	**1.114** (90.2%)
Michael Jackson	0.869 (56.8%)	6.527 (98.9%)
The Supremes	0.762 (53.6%)	5.068 (98.2%)
Madonna	**3.901 (87.1%)**	1.344 (91.3%)
Whitney Houston	2.010 (76.5%)	2.369 (94.7%)
Stevie Wonder	0.247 (29.0%)	2.297 (94.5%)
Janet Jackson	1.162 (64.2%)	1.723 (92.8%)

Table 7: Top artists, ordered by number-one singles. Raw values and percentiles are shown.

The all-time top writers of number-one hits have influence scores close to 1, with more successful writers having slightly lower influence scores (Table 6). This result mirrors that found for central genres and suggests that these writers were well connected in a lyrical culture that they both contributed to and drew from. The exception is Max Martin, whose lower influence score perhaps reflects the less quotable nonsense phrases he often uses in songs [12]. Martin is also the least self-referential of the top writers, which might be explained by noting that he writes for a variety of artists with different styles, whereas the other writers are most famous for writing for themselves.

Table 7 shows the top 10 artists by number-one hits. In contrast to writer position, artist position does not seem strongly affected by influence score. However, all top artists fall into the top decile (10%) of self-reference. Also, female artists in the list have much higher influence scores than males, with the exception of The Supremes and Rihanna. We note also that Mariah Carey has a much higher influence score as an artist than as a writer. Further analysis of this phenomenon is complicated by the fact that many people use a pseudonym as an artist but not as a writer.

4. DISCUSSION

We explored topologies of genre, artist, and songwriter influence networks constructed from links between trigrams over time. Through edge reduction and strongly connected component analyses, we showed that all three graphs are highly centralized around a large component with robust links. We confirmed this organization with an MDS visualization of the genre graph based on mutual influence. Alternative methods of edge reduction revealed separate components, but primarily along language differences. We found that the best predictor of a genre's centrality to the influence network was the degree to which it referenced itself, and that the network especially centered around three popular and self-referential genres: Jazz, Pop, and Rock.

Our current metrics are useful building blocks for studying relationships between genres, artists, and writers. The centrality of our influence networks supplements earlier findings showing clustering between some genres and isolation of others [6, 9]. However, our data do not produce the Folk, Country, and Blues cluster observed by Fell and Sporleder [6]; rather, we find these genres have similar influences but do not draw significant influence from each other. Fell and Sporleder also found Rap and Metal to be lyrically isolated when analyzed with a bag-of-words model [6]. Our trigram analysis shows Rap and Metal to be well connected in the network, though their self-reference scores are low compared to other popular genres. Furthermore, the notion of lyrical novelty [5] can be approximated with influence, as the influence network incorporates novelty into the edge weights with tf-idf; cliché [14] can be understood as the inverse of novelty. Overall, our analyses do not suggest the same degree of genre segmentation suggested in prior studies. We conclude that significant differences between genres may not occur at the phrase level, but instead arise from key vocabulary differences [6, 9] as well as musical and sociocultural factors [2, 15].

There are several potential areas for improvement in the present study. First, a phrase shared across songs does not necessarily signify direct influence. Next, albums were occasionally labeled with incorrect years, which would impact the temporal dimension of our networks. We used primary iTunes album genre for our analysis, but acknowledge that such labels may not adequately characterize the songs, especially for non-Western music or for songs that conceivably belong to more than one genre. Also, the decrease in connection robustness from cleaning may not be uniform across genres. In particular, the present results could be refined by analyzing English lyrics only or by including stopwords and cleaning rules for other languages. The cleaning procedure we followed [5] may benefit from stemming. Finally, the decision to ignore phrases spanning lines of lyrics reflects the organization of phrases for most genres, but may have broken up phrases from genres with more complicated lyrics (such as Rap).

Our findings highlight exciting possibilities for future research. Recall that at one point in our component analyses, Jazz was excluded from the central component comprising Pop and Rock—not because it was less influential, but because it drew much less influence from Pop and Rock than they drew from it, leaving no path back to Jazz from Pop or Rock when only the highest-magnitude edges remained in the graph. Future analyses could examine whether the edge reduction component analysis aligns more closely with self-reference when graphs are treated as undirected. Future work could also find cliques, which would be more robustly interconnected than strongly connected components. Assessing changes in network structure when higher-order n-grams are used is another topic that can be explored in future research. Finally, future studies could examine further the notion of robustness of connection; differences in influence when people act as artists versus writers; trends that emerge when people are grouped by gender, race, or geolocation; network topologies when rare genres are grouped into categories; and could contribute visualizations to help understand the structure of artist and writer networks.

Acknowledgements: The authors thank Will Mills from LyricFind for facilitating access to the lyrics data; and Casey Baker and Ge Wang for their helpful feedback.

5. REFERENCES

[1] J. Atherton and B. Kaneshiro. Lyrical influence networks dataset (LIND). In *Stanford Digital Repository*, 2016. http://purl.stanford.edu/zy061bp9773.

[2] N. J. Bryan and G. Wang. Musical influence network analysis and rank of sample-based music. In *Proceedings of the 12th International Society for Music Information Retrieval Conference*, pages 329–334, 2011.

[3] N. Collins. Computational analysis of musical influence: A musicological case study using MIR tools. In *Proceedings of the 11th International Society for Music Information Retrieval Conference*, pages 177–182, 2010.

[4] C. N. DeWall, R. S. Pond Jr, W. K. Campbell, and J. M. Twenge. Tuning in to psychological change: Linguistic markers of psychological traits and emotions over time in popular US song lyrics. *Psychology of Aesthetics, Creativity, and the Arts*, 5(3):200–207, 2011.

[5] R. J. Ellis, Z. Xing, J. Fang, and Y. Wang. Quantifying lexical novelty in song lyrics. In *Proceedings of the 16th International Society for Music Information Retrieval Conference*, pages 694–700, 2015.

[6] M. Fell and C. Sporleder. Lyrics-based analysis and classification of music. In *Proceedings of the International Conference on Computational Linguistics*, pages 620–631, 2014.

[7] H. Fujihara, M. Goto, and J. Ogata. Hyperlinking lyrics: A method for creating hyperlinks between phrases in song lyrics. In *Proceedings of the 9th International Conference on Music Information Retrieval*, pages 281–286, 2008.

[8] C. Gunaratna, E. Stoner, and R. Menezes. Using network sciences to rank musicians and composers in Brazilian popular music. In *Proceedings of the 12th International Society for Music Information Retrieval Conference*, pages 441–446, 2011.

[9] B. Logan, A. Kositsky, and P. Moreno. Semantic analysis of song lyrics. In *IEEE International Conference on Multimedia and Expo*, volume 2, pages 827–830, 2004.

[10] J. P. G. Mahedero, A. Martínez, P. Cano, M. Koppenberger, and F. Gouyon. Natural language processing of lyrics. In *Proceedings of the 13th annual ACM International Conference on Multimedia*, pages 475–478, 2005.

[11] G. Salton and M. J. McGill. *Introduction to modern information retrieval*. McGraw-Hill, Inc., New York, 1986.

[12] J. Seabrook. Blank space: What kind of genius is Max Martin? *New York Times*, Sep. 30, 2015.

[13] R. N. Shepard. Multidimensional scaling, tree-fitting, and clustering. *Science*, 210(4468):390–398, 1980.

[14] A. Smith, C. Zee, and A. Uitdenbogerd. In your eyes: Identifying clichés in song lyrics. In *Australasian Language Technology Workshop*, pages 88–96, 2012.

[15] R. D. Smith. The network of collaboration among rappers and its community structure. *Journal of Statistical Mechanics: Theory and Experiment*, 2006(02):2–15, 2006.

[16] M. van Zaanen and P. Kanters. Automatic mood classification using TF*IDF based on lyrics. In *Proceedings of the 11th International Society for Music Information Retrieval Conference*, pages 75–80, 2010.

[17] X. Wang, X. Chen, D. Yang, and Y. Wu. Music emotion classification of Chinese songs based on lyrics using TF*IDF and rhyme. In *Proceedings of the 12th International Society for Music Information Retrieval Conference*, pages 765–770, 2011.