

# Musically Meaningful or Just Noise? An Analysis of On-line Artist Networks

Kurt Jacobson<sup>1</sup> and Mark Sandler<sup>1</sup>

Centre for Digital Music, Queen Mary University  
{kurt.jacobson, mark.sandler}@elec.qmul.ac.uk  
<http://www.elec.qmul.ac.uk/digitalmusic/>

**Abstract.** A sample of the Myspace social network is examined. Using methods from complex network theory, we show empirically that artists tend to form on-line social connections with artists of the same genre. This motivates the use of on-line social networks as data resources for musicology and music information retrieval.

## 1 Introduction

Music is an inherently social phenomenon. Of course music can be practiced in solitude, providing the individual with personal satisfaction and even profound inner-vision. But it is the social element - performance, participation, dance - that makes music a truly universal cultural foundation.

Musicologists have long recognized that in discussing music, it is important to consider its context - who created the music? where? why? One important aspect of this is the social context - who were the artist's contemporaries? her friends? her influences? [1].

Today, online social networking websites can provide a concise cross section of this information in an easily accessible format. Myspace has become the de-facto standard for web-based music artist promotion. Although exact figures are not made public, recent blogosphere chatter suggests there are around 7 million artist pages on Myspace [2].

Artists ranging from bedroom electronica amateurs to multi-platinum megastars publish Myspace pages. These Myspace artist pages typically include some media - usually streaming audio, video, or both - and a list of "friends" specifying social connections. This combination of media and a user-specified social network provides a unique data set that is unprecedented in both scope and scale. We examine a sample of this on-line social network of artists. Comparing the network topology with a collection of genre tags associated with each artist, we show there is a tendency towards *homophily* - artists are more likely to be friends with artists of the same genre. This suggests such networks are meaningful in the context of several music-related studies including music information retrieval and social musicology.

We begin with a review of some concepts from the complex networks literature and their applications to musician and listener networks. In section 3 we

describe the methods used to gather our sample of the Myspace artist network. This is followed by our empirical results in section 4 and a discussion of the results and suggestions for applications and future work in section 5.

## 2 Complex Networks

Complex network theory deals with the structure of relationships in complex systems. Using the tools of graph theory and statistical mechanics, physicists have developed models and metrics for describing a diverse set of real-world networks - including social networks, academic citation networks, biological protein networks, and the World-Wide Web. All these networks exhibit several unifying characteristics such as small worldness, scale-free degree distributions, and community structure [3],[4]. Let us briefly discuss some definitions and concepts that will be used in this work.

### 2.1 Network Properties

A given network  $G$  is described by a set of *nodes*  $N$  connected by a set of *edges*  $E$ . Each edge is defined by the pair of nodes it connects  $(i, j)$ . If the edges imply directionality,  $(i, j) \neq (j, i)$ , the network is a *directed network*. Otherwise, it is an *undirected network*. The number of edges incident to the a node  $i$  is the *degree*  $k_i$ . In a directed network there will be an *indegree*  $k_i^{in}$  and an *outdegree*  $k_i^{out}$  corresponding to the number of edges pointing into the node and away from the node respectively.

**Degree distribution** The *degree distribution*  $P(k)$  is the proportion of nodes that have a degree  $k$ . The shape of the degree distribution is an important metric for classifying a network - “scale-free networks” have a power-law distribution [4] while “random networks” have a Poisson distribution. The scale-free degree distribution is a property common to many real-world networks. Conceptually, a scale-free distribution indicates the presence of a few very-popular *hubs* that tend to attract more links as the network evolves [3], [4].

**Average shortest path** Two nodes  $i$  and  $j$  are connected if a path exists between them following the edges in the network. The path from  $i$  to  $j$  may not be unique. The *geodesic path*  $d_{ij}$  is the shortest path distance from  $i$  to  $j$  in number of edges traversed. For the entire network, the average shortest path or mean geodesic distance is  $l$ .

$$l = \frac{1}{\frac{1}{2}n(n+1)} \sum_{i \geq j} d_{ij} \quad (1)$$

where  $d_{ij}$  is the geodesic distance from node  $i$  to node  $j$  and  $n$  is the total number of nodes in the network. In a “small-world network” the mean geodesic distance is small relative to the number of nodes in the network [4], [3]. The largest geodesic distance in a network is known as the *diameter*.

**Transitivity** The transitivity or clustering coefficient estimates the probability that two neighboring nodes of a given node are neighbors themselves. In the terms of social networks, the friend of your friend is also likely to be your friend. In terms of network topology, transitivity means a heightened number of triangles exist - sets of three nodes that are each connected to each other. For a given undirected unweighted network the transitivity is defined as

$$C = \frac{3N_{\Delta}}{N_3} \quad (2)$$

where  $N_{\Delta}$  is the number of triangles in the network and  $N_3$  is the number of connected triples. A connected triple is a set of three nodes where each node can be reached from every other node.

## 2.2 Networks and Music

Quite naturally, networks of musicians have been studied in the context of complex network theory - typically viewing the artists as nodes in the network and using either collaboration, influence, or similarity to define network edges. These networks of musicians exhibit many of the properties expected in social networks [5], [6], [7]. Networks of music listeners and bipartite networks of listeners and artists have also been studied [8], [9].

## 3 Sampling Myspace

The Myspace social network presents a variety of challenges. For one, the massive size prohibits analyzing the graph in its entirety, even when considering only the artist pages. Therefore we sample a small yet sufficiently large portion of the network as described in section 3.2. Also, the Myspace social network is filled with noisy data - plagued by spammers and orphaned accounts. We limit the scope of our sampling in a way that minimizes this noise. And finally, there currently is no interface for easily collecting the network data from Myspace. Our data is collected using web crawling and screen scraping techniques.

### 3.1 Artist Pages

It is important to note we are only concerned with a subset of the Myspace social network - the Myspace *artist* network. Myspace artist pages are different from standard Myspace pages in that they include a distinct audio player application. We use the presence or absence of this player to determine whether or not a given page is an artist page.

A Myspace page will most often include a top friends list. This is a hyper-linked list of other Myspace accounts explicitly specified by the user. The top friends list is limited in length with a maximum length of 40 friends (the default length is 16 friends). In constructing our sampled artist network, we use the top friends list to create a set of directed edges between artists. Only top friends

who also have artist pages are added to the sampled network; standard Myspace pages are ignored. We also ignore the remainder of the friends list (i.e. friends that are not specified by the user as top friends), assuming these relationships are not as relevant. This reduces the amount of noise in the sampled network but also artificially limits the outdegree of each node. Our sampling method is based on the assumption that artists specified as top friends have some meaningful musical connection for the user – whether through collaboration, stylistic similarity, friendship, or artistic influence.

Each Myspace artist page includes between zero and three genre tags. The artist selects from a list of 119 genres specified by Myspace. In our sample set, around 2.6% of artists specified no genre tags.

### 3.2 Snowball Sampling

There are several network sampling methods; however, for the Myspace artist network, snowball sampling is the only appropriate method [10], [11]. In this method, a first seed node (artist page) is included in the sample. Then the seed node’s neighbors (top friends) are included in the sample. Then the neighbors’ neighbors. This breadth-first sampling is continued until a particular sampling ratio is achieved. Here, we go through 6 iterations of sampling to collect 15,478 nodes. If the size of the Myspace artist network is around 7 million, then this is close to the 0.25% sampling ratio suggested for accurate degree distribution estimation in sampled networks [12].

With snowball sampling there is a tendency to over-sample hubs because they have many links and are easily picked up early in the breadth-first sampling. This property would reduce the degree distribution exponent and produce a heavier tail but preserve the power-law nature of the network [11].

## 4 Empirical Results

The Myspace artist network sample conforms in many respects to the topologies expected in social networks. The network statistics for our sample are summarized in Table 1.

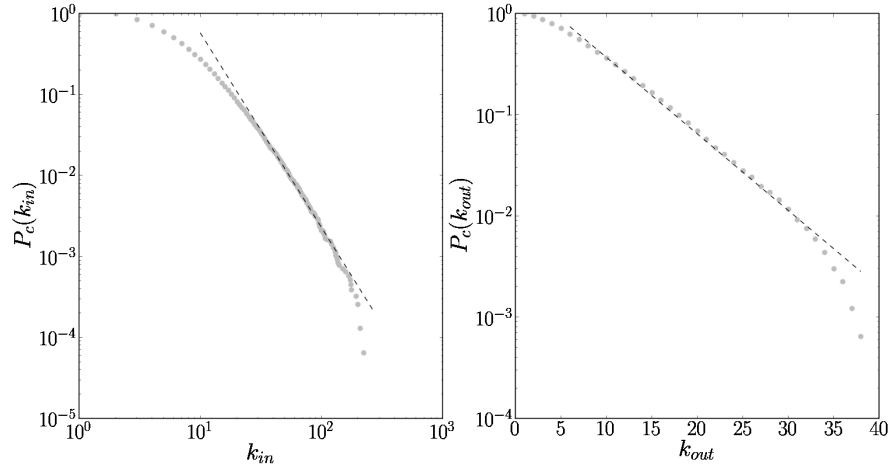
The values for the mean geodesic distance  $l$  and the diameter  $d_{max}$  are relatively small and suggest a small-world network - any node can be reached from any other node in a small number of steps. The value of the clustering coefficient  $C = 0.219$  indicates a high level community structure - an equivalent random network would have a transitivity of  $C_r = \frac{\langle k \rangle}{n} = 7.26 \cdot 10^{-4}$ . Both of these properties are commonly reported in social networks [4], [3]. While our sample size is large enough for estimating the degree distributions of the entire network, it is insufficient to assume the  $C$ ,  $l$ , and  $d_{max}$  values hold true for the entire Myspace artist network [12].

The cumulative degree distributions for the network sample are plotted in Fig. 1. Both the in-degree and out-degree distributions are plotted. Notice the in-degree distribution is plotted on a log-log scale. For moderate degree values

**Table 1.** The network statistics for the Myspace artist network sample where  $n$  is the number of nodes,  $m$  is the number of edges,  $\langle k \rangle$  is the average degree,  $l$  is the mean geodesic distance,  $d_{max}$  is the diameter, and  $C$  is the clustering coefficient. The clustering coefficient is undefined for directed networks

	$n$	$m$	$\langle k \rangle$	$l$	$d_{max}$	$C$
undirected	15,478	91,326	11.801	4.479	9	.219
directed	15,478	120,487	15.569	6.426	16	-

( $35 < k_{in} < 200$ ), the in-degree distribution follows a power-law decay. The power-law fit breaks down for high and low values of  $k_{in}$ . Similar “broad-scale” degree distributions have been reported for citation networks and movie actor networks [13].



**Fig. 1.** The cumulative degree distributions for (a) the in degree plotted on a log-log scale and (b) the out degree plotted on a log-linear scale

The out-degree distribution is plotted on a log-linear scale and follows a distinct exponential decay. Such distributions have been reported in a variety of networks including some music recommendation networks [6]. This distribution is most likely related to the out degree limit imposed by Myspace which restricts the maximum number of top friends ( $k_{out} \leq 40$ ).

#### 4.1 Assortativity with Respect to Genre

We are most interested in the mixing patterns found in our network sample. Do artists tend to form friendships with artists of the same genre? To answer this question we look at the homophily or *assortative mixing* with respect to genre tags for our network. If a network exhibits assortative mixing, nodes tend to form links with nodes of the same type, or in our case, artists tend to form friendships with other artists of the same genre. The assortativity coefficient for our network sample is calculated following [4].

Let  $E$  be an  $N \times N$  matrix with elements  $E_{\gamma_i \gamma_j}$ . For genre labels  $\gamma_1, \gamma_2, \dots, \gamma_N$ , let  $E_{\gamma_i \gamma_j}$  be the number of edges in a network that connect vertices of genre  $\gamma_i$  and  $\gamma_j$ . The normalized mixing matrix is defined as

$$e = \frac{E}{\|E\|} \quad (3)$$

where  $\|x\|$  means the sum of all elements in the matrix  $x$ . The elements  $e_{\gamma_i \gamma_j}$  measure the fraction of edges that fall between nodes of genre  $\gamma_i$  and  $\gamma_j$ . The assortativity coefficient  $r$  is then defined as

$$r = \frac{\text{Tr}(e) - \|e^2\|}{1 - \|e^2\|} \quad (4)$$

This quantity will be 0 in a randomly mixed network, 1 in a perfectly assortative network, and negative for a disassortative network where nodes only connect with nodes of different types. However, this assortativity calculation assumes each node can only be of one type. In our sample, each artist is associated with between 0 and 3 genre types. For simplicity, we truncate the list of genre labels for each artist taking only the first label. Artists with no genre label are excluded from the calculation. Using this approach the assortativity coefficient with respect to genre for our sample is  $r = 0.350$ . Of course this calculation assumes an overly strict definition of genre type, excluding all but one genre label. Also, this calculation also makes no allowance for genre types that are conceptually very similar such as ‘‘Hip-Hop’’ and ‘‘Rap.’’ Therefore we can assume this value of  $r$  underestimates the actual level of assortative mixing in the network sample. A classic example of assortative mixing in real-world networks is the network of sexual partnerships which shows strong assortative mixing by race with  $r = 0.621$ .

Another approach is to preserve the entire genre label set and accept a weak definition for genre type - two artists are of the same genre type if they have one or more label in common. Again, artists with no genre label are excluded. Then for each artist pair, if one or more genre labels are shared, the first shared label  $\gamma_i$  is selected and one count is added at  $E_{\gamma_i \gamma_i}$ . If no genre labels are shared, the first genre label for each artist is selected and one count is added at  $E_{\gamma_i \gamma_j}$ . In this way, only one count is added for each node pair in the network.

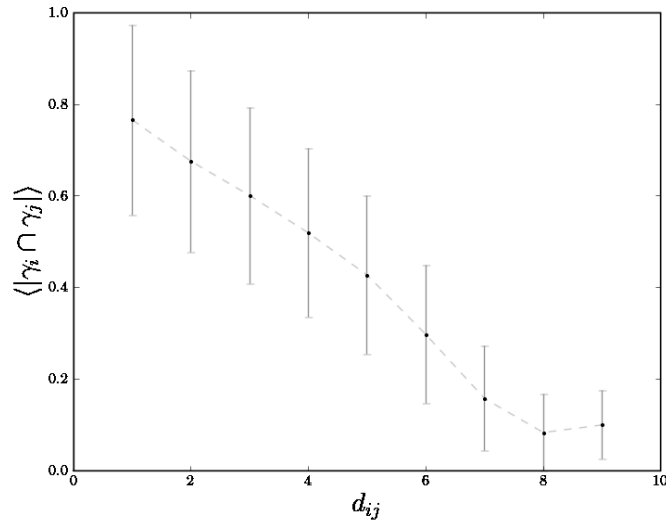
Using the weak definition of genre type, the assortativity coefficient for our sample with respect to genre is  $r = 0.778$ . Of course this is allowing each node to be recast as a different genre types to facilitate matches, somewhat inflating

the assortativity coefficient. However, this reflects the reality of artist-genre associations where a single artist is often associated with several genres. And this calculation still makes no allowance for genres that are conceptually very similar, underestimating the assortative mixing in this sense. Still, the calculations show that artists prefer to maintain friendship links with other artists in the same genre in the Myspace artist network.

## 4.2 Distance and Genre

Let us also examine the number of shared genre labels as a function of geodesic distance between artists. That is, do artist that are closer to each other in the network have more genre labels in common?

For this analysis let us take a undirected view of the graph. Each edge is considered bi-directional, that is  $(i, j) = (j, i)$ , and if a reflexive pair of edges existed in the directed graph, only one bi-directional edge exists in the undirected graph. This reduces the edge count by about 24% (see Table 1).



**Fig. 2.** The geodesic distance  $d_{ij}$  vs. the average number of shared genre labels between nodes  $\langle \|\gamma_i \cap \gamma_j\| \rangle$

For each node pair  $i$  and  $j$  in the undirected network, we calculate the geodesic distance  $d_{ij}$ . Then, we count how many genre labels are shared between the nodes  $\|\gamma_i \cap \gamma_j\|$ , where node  $i$  has a set of genre labels  $\gamma_i$  and node  $j$  has labels  $\gamma_j$ . For each value of  $d_{ij}$  we plot the average value of  $\|\gamma_i \cap \gamma_j\|$ . This shows how the number of genre labels shared between node pairs is related to

geodesic distance. The results are shown in Fig. 2. The averages are plotted with standard deviation error bars. There is a clear decrease in the number of shared genre labels as the geodesic distance increases. However we do see an increase at  $d_{ij} = 9$ . It should be noted that in our network sample a geodesic distance of 9 is extremely rare. For the  $1.2 \cdot 10^8$  node pairs in the undirected network, only 20 have  $d_{ij} = 9$ .

## 5 Discussion

Studying the sample of the Myspace artist network indicates that its structure is quite relevant to music. There is evidence of high assortativity with respect to genre. In an on-line social network, artists tend to form friendships with artists of the same genre.

The plot in Fig. 2 indicates a clear relationship between genre similarity and geodesic distance. In the on-line social network, artists tend to be more closely connected to other artists with similar genre labels.

These findings indicate that the structure of the Myspace artist network reflects the genre structures traditionally used to classify music. This result is not surprising, but it is significant - motivating new network-based genre-fication approaches - grouping artists based on network topologies. A similar approach is taken in [9], although the focus is on listener-generated data. Musicological metrics could be developed for quantifying the level of influence of a particular artist based on network structure.

Applying network community-finding algorithms to this data set could yield some interesting results. For example genre distributions within a community could be used to evaluate different community-finding algorithms. A similar approach is used in [8] but using networks of listeners instead of artists. Identifying communities of artists could be helpful in music recommendation applications - if a user expresses a preference for a particular artist, recommend other artists in that community. Also, analysis of these artist community structures could yield some interesting insights in social musicology studies.

In another analysis of this data set, we examine the relationship between audio-based similarity and the network structure [14]. We find that artist pairs with a shorter geodesic distance in the network tend to produce more similar music from a signal analysis perspective.

Analysis of this data set has shown, despite the unregulated nature of the Myspace artist network, its structure reflects the top-down genre structures traditionally used to classify music. This suggests such social network data could be of interest in a variety of contexts including music recommendation, music information retrieval, and musicology.

## 6 Acknowledgements

This work is supported as a part of the OMRAS2 project, EPSRC grants EP/E02274X/1 and EP/E017614/1.

## References

1. J. Shepherd. Sociology of music. [Online]. Available: <http://www.groovemusic.com>
2. [Online]. Available: <http://scottelkin.com/archive/2007/05/11/Myspace-Statistics.aspx>
3. L. da F. Costa, F. A. Rodrigues, G. Travieso, and P. R. V. Boas, "Characterization of complex networks: a survey of measurements," *Advanced Physics*, vol. 56, pp. 167–242, 2007.
4. M. E. J. Newman, "The structure and function of complex networks," *SIAM Review*, vol. 45, pp. 167–256, 2003.
5. P. Gleiser and L. Danon, "Community structure in jazz," *Advances in Complex Systems*, vol. 6, no. 4, pp. 565–573, 2003.
6. P. Cano, O. Celma, and M. Koppenberger, "Topology of music recommendation networks," *Chaos*, vol. 16, 2006.
7. J. Park, O. Celma, M. Koppenberger, P. Cano, and J. M. Buldu, "The social network of contemporary popular musicians," *Physics and Society*, 2006.
8. A. Anglade, M. Tiemann, and F. Vignoli, "Virtual communities for creating shared music channels," in *Proc. of Int. Conference on Music Information Retrieval*, 2007.
9. R. Lambiotte and M. Ausloos, "On the genre-fication of music: a percolation approach," *Eur. Phys. J.*, vol. 50, pp. 183–188, 2006.
10. Y.-Y. Ahn, S. Han, H. Kwak, S. Moon, and H. Jeong, "Analysis of topological characteristics of huge online social networking services," in *Proceedings of the 16th International Conference on World Wide Web*. Alberta, Canada: IW3C2, May 2007.
11. S. H. Lee, P.-J. Kim, and H. Jeong, "Statistical properties of sampled networks," *Physical Review E*, vol. 73, pp. 102–109, January 2006.
12. H. Kwak, S. Han, Y.-Y. Ahn, S. Moon, and H. Jeong, "Impact of snowball sampling ratios on network characteristics estimation: A case study of cyworld," KAIST, Tech. Rep. CS/TR-2006-262, November 2006.
13. L. A. N. Amaral, A. Scala, M. Barthélemy, and H. E. Stanley, "Classes of small-world networks," in *Proceeding of the National Academy of Sciences*, 2000.
14. 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," January 2008, submitted for publication.