USING AUDIO ANALYSIS AND NETWORK STRUCTURE TO IDENTIFY COMMUNITIES IN ON-LINE SOCIAL NETWORKS OF ARTISTS

KURT JACOBSON, MARK SANDLER
Centre for Digital Music
Queen Mary, University of London
kurt.jacobson@elec.qmul.ac.uk

BEN FIELDS
Goldsmiths Digital Studios
Goldsmiths, University of London
b.fields@gold.ac.uk

ABSTRACT

Community detection methods from complex network theory are applied to a subset of the Myspace artist network to identify groups of similar artists. Methods based on the greedy optimization of modularity and random walks are used. In a second iteration, inter-artist audio-based similarity scores are used as input to enhance these community detection methods. The resulting community structures are evaluated using a collection of artist-assigned genre tags. Evidence suggesting the Myspace artist network structure is closely related to musical genre is presented and a Semantic Web service for accessing this structure is described.

1 INTRODUCTION

The dramatic increase in popularity of online social networking has led hundreds of millions of individuals to publish personal information on the Web. Music artists are no exception. Myspace [1] 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 well over 7 million artist pages [2] on Myspace. Myspace artist pages typically include some streaming audio 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.

However, the Myspace network is the result of hundreds of millions individuals interacting in a virtually unregulated fashion. Can this crowd-sourced tangle of social networking ties provide insights into the dynamics of popular music? Does the structure of the Myspace artist network have any relevance to music-related studies such as music recommendation or musicology?

In an effort to answer these questions, we identify communities of artists based on the Myspace network topology and attempt to relate these community structures to musical genre. To this end, we examine a sample of the Myspace social network of artists. First we review some previous work on the topics of artist networks, audio-based music analysis, and complex network community identification. We then describe our methodology including our network sampling method in Section 3.1 and our community detection approaches in Section 3.2. In Section 3.3 we describe the concept of *genre entropy* - a metric for evaluating the relevance of these community structures to music. Finally, we include a discussion of the results, suggestions for future work, and describe a Semantic Web service that can be used to access some of the data in a structured format.

[1] `http://myspace.com`

[2] `http://scottelkin.com/archive/2007/05/11/Myspace-Statistics.aspx`
∼25 million songs, ∼3.5 songs/artist, ∼7 million artists

2 BACKGROUND

2.1 Complex Networks

Complex network theory uses the tools of graph theory and statistical mechanics to deal with the structure of relationships in complex systems. A network is defined as a graph $G = (N, E)$ where N is a set of *nodes* connected by a set of *edges* E. We will refer to the number of nodes as n and the number of edges as m. The network can also be defined in terms of the *adjacency matrix* $G = A$ where the elements of A are

$$A_{ij} = \begin{cases} 1 & \text{if nodes } i \text{ and } j \text{ are connected,} \\ 0 & \text{otherwise.} \end{cases} \tag{1}$$

In this work, we restrict our analysis to the *undirected* case where edges are not considered directional and A is a symmetric matrix. For a summary of recent developments in complex networks see [7, 17].

2.2 Music Networks

Networks of musicians have been studied in the context of complex network theory - viewing the artists as nodes in the network and using either collaboration, influence, or some measure of similarity to define network edges [4, 5, 9, 19]. However the networks studied are generally constructed based on expert opinions (e.g. AllMusicGuide [3]) or proprietary

[3] `http://www.allmusic.com/`

algorithms based on user listening habits (e.g. Last.fm [4]). The Myspace artist network is unique in that the edges - the "friend" connections - are specified by the artists themselves. This makes the Myspace artist network a true social network. It has been shown that significantly different network topologies result from different approaches to artist network construction [4]. Since the Myspace artist network is of unique construction - owing its structure to the decisions and interactions of millions of individuals - we are motivated to analyze its topology and explore how this network structure relates to music.

It should be noted that networks of music listeners and bipartite networks of listeners and artists have also been studied [2, 13]. While such studies are highly interesting in the context of music recommendation, and while the Myspace network could potentially provide interesting data on networks of listeners, we restrict our current investigation to the Myspace artist network.

Previous analysis of the Myspace social network (including artists and non-artists) suggests that it conforms in many respects to the topologies commonly reported in social networks - having a power-law degree distribution and a small average distance between nodes [1]. Previous analysis of the Myspace *artist* network sample used in this work shows a multi-scaling degree distribution, a small average distance between nodes, and strong assortative mixing with respect to genre [11].

2.3 Community Detection

Recently, there has been a significant amount of interest in algorithms for detecting community structures in networks. These algorithms are meant to find dense subgraphs (communities) in a larger sparse graph. More formally, the goal is to find a partition $\mathcal{P} = \{C_1, \ldots, C_c\}$ of the nodes in graph G such that the proportion of edges inside C_k is high compared to the proportion of edges between C_k and other partitions.

Because our network sample is moderately large, we restrict our analysis to use more scalable community detection algorithms. We make use of the greedy modularity optimization algorithm [6] and the walktrap algorithm [20]. These algorithms are described in detail in Section 3.2.

2.4 Signal-based Music Analysis

A variety of methods have been developed for signal-based music analysis, characterizing a music signal by its timbre, harmony, rhythm, or structure. One of the most widely used methods is the application of Mel-frequency cepstral coefficients (MFCC) to the modeling of timbre [15]. In combination with various statistical techniques, MFCCs have been successfully applied to music similarity and genre classification tasks [18, 16, 3, 10]. A common approach for computing timbre-based similarity between two songs or collections of songs creates Gaussian mixtures models (GMM) describing the MFCCs and comparing the GMMs using a statistical distance measure. Often the earth mover's distance (EMD), a technique first used in computer vision, is the distance measure used for this purpose [21]. The EMD algorithm finds the minimum work required to transform one distribution into another. We use a set of inter-artist EMD values as a means of enhancing our community detection methods as described in Section 3.2.3.

3 METHODOLOGY

We will review our methodology beginning with a description of our network sampling method in Section 3.1. We then describe the various community detection approaches applied to the network in Section 3.2 and how we incorporate audio-based measures. Finally, we describe our metric for evaluating the relevance of the Myspace artist network structure with respect to musical genre in Section 3.3.

3.1 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 3.1.2. 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.

3.1.1 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 always include a top friends list. This is a hyperlinked 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

[4] `http://last.fm`

sampled network but also artificially limits the outdegree of each node. This approach 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. We include this information in our data set.

The audio files associated with each artist page in the sampled network are also collected for feature extraction as described in Section 3.2.3.

3.1.2 Snowball Sampling

For the Myspace artist network, snowball sampling is the most appropriate method [1]. Alternative methods such as random edge sampling and random node sampling would result in many small disconnected components and not provide any insight to the structure of the entire network [14]. In snowball sampling, 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. We randomly select one seed node [5] and perform 6 levels of sampling - such that in an undirected view of the network, no artist can have a geodesic distance greater than 6 with respect to the seed artist - 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 in [12].

3.1.3 Conversion to Undirected Graph

With the sampling method described above, the edges in our Myspace artist network are directional. If j is a top friend of i, this does not mean i is a top friend of j ($(i, j) \neq (j, i)$). However, many community detection algorithms operate on *undirected* graphs where $(i, j) = (j, i)$. For this reason we convert our directed graph to an undirected graph. Where a single directed edge exists it becomes undirected and where a reflexive pair of directed edges exist a single undirected edge replaces both edges. This process reduces the edge count from $120,487$ to $91,326$.

3.2 Community Detection

We apply two community detection algorithms to our network sample - the greedy optimization of modularity [6] and the walktrap algorithm [20]. Both of these algorithms are reasonably efficient and both algorithms can be easily adapted to incorporate audio-based similarity measures.

[5] our randomly selected artist is French rapper Karna Zoo http://www.myspace.com/karnazoo

3.2.1 Greedy Modularity Optimization

Modularity is a network property that measures the appropriateness of a network division with respect to network structure. Modularity can be defined in several different ways [7]. In general, modularity Q is defined as the number of edges within communities minus the expected number of such edges. Let A_{ij} be an element of the network's adjacency matrix and suppose the nodes are divided into communities such that node i belongs to community C_i. We define modularity Q as the fraction of edges within communities minus the expected value of the same quantity for a random network. Then Q can be calculated as follows:

$$Q = \frac{1}{2m} \sum_{ij} \left[A_{ij} - \frac{d_i d_j}{2m} \right] \delta_{C_i C_j} \qquad (2)$$

where the $\delta_{C_i C_j}$ function is 1 if $C_i = C_j$ and 0 otherwise, m is the number of edges in the graph, and d_i is the *degree* of node i - that is, the number of edges incident on node i. The sum of the term $\frac{d_i d_j}{2m}$ over all node pairs in a community represents the expected fraction of edges within that community in an equivalent random network where node degree values are preserved.

If we consider Q to be a benefit function we wish to maximize, we can then use an agglomerative approach to detect communities - starting with a community for each node such that the number of partitions $|\mathcal{P}| = n$ and building communities by amalgamation. The algorithm is greedy, finding the changes in Q that would result from the merge of each pair of communities, choosing the merge that results in the largest increase of Q, and then performing the corresponding community merge. It can be proven that if no community merge will increase Q the algorithm can be stopped because no further modularity optimization is possible [6]. Using efficient data structures based on sparse matrices, this algorithm can be performed in time $\mathcal{O}(m \log n)$.

3.2.2 Random Walk: Walktrap

The walktrap algorithm uses random walks on G to identify communities. Because communities are more densely connected, a random walk will tend to be 'trapped' inside a community - hence the name "walktrap".

At each time step in the random walk, the walker is at a node and moves to another node chosen randomly and uniformly from its neighbors. The sequence of visited nodes is a *Markov chain* where the states are the nodes of G. At each step the transition probability from node i to node j is $P_{ij} = \frac{A_{ij}}{d_i}$ which is an element of the transition matrix P for the random walk. We can also write $P = D^{-1}A$ where D is the diagonal matrix of the degrees ($\forall i, D_{ii} = d_i$ and $D_{ij} = 0$ where $i \neq j$).

The random walk process is driven by powers of P: the probability of going from i to j in a random walk of length

t is $(P^t)_{ij}$ which we will denote simply as P^t_{ij}. All of the transition probabilities related to node i are contained in the i^{th} row of P^t denoted as $P^t_{i\bullet}$. We then define an inter-node distance measure:

$$r_{ij} = \sqrt{\sum_{k=1}^{n} \frac{(P^t_{ik} - P^t_{jk})^2}{d_k}} = \|D^{-\frac{1}{2}} P^t_{i\bullet} - D^{-\frac{1}{2}} P^t_{j\bullet}\| \quad (3)$$

where $\|.\|$ is the Euclidean norm of $\Re^n$. This distance can also be generalized as a distance between communities: $r_{C_i C_j}$ or as a distance between a community and a node: $r_{C_i j}$.

We then use this distance measure in our algorithm. Again, the algorithm uses an agglomerative approach, beginning with one partition for each node ($|\mathcal{P}| = n$). We first compute the distances for all adjacent communities (or nodes in the first step). At each step k, two communities are chosen based on the minimization of the mean σ_k of the squared distances between each node and its community.

$$\sigma_k = \frac{1}{n} \sum_{C_i \in \mathcal{P}_k} \sum_{i \in C_i} r^2_{iC_i} \quad (4)$$

Direct calculation of this quantity is known to be NP-hard, so instead we calculate the variations $\Delta\sigma_k$. Because the algorithm uses a Euclidean distance, we can efficiently calculate these variations as

$$\Delta\sigma(C_1, C_2) = \frac{1}{n} \frac{|C_1||C_2|}{|C_1| + |C_2|} r^2_{C_1 C_2} \quad (5)$$

The community merge that results in the lowest $\Delta\sigma$ is performed. We then update our transition probability matrix

$$P^t_{(C_1 \cup C_2)\bullet} = \frac{|C_1| P^t_{C_1\bullet} + |C_2| P^t_{C_2\bullet}}{|C_1| + |C_2|} \quad (6)$$

and repeat the process updating the values of r and $\Delta\sigma$ then performing the next merge. After $n-1$ steps, we get one partition that includes all the nodes of the network $\mathcal{P}_n = \{N\}$. The algorithm creates a sequence of partitions $(\mathcal{P}_k)_{1 \le k \le n}$. Finally, we use modularity to select the best partition of the network, calculating $Q_{\mathcal{P}_k}$ for each partition and selecting the partition that maximizes modularity.

Because the value of t is generally low (we use $t = 4$), this community detection algorithm is quite scalable. For most real-world networks, where the graph is sparse, this algorithm runs in time $\mathcal{O}(n^2 \log n)$ [20].

3.2.3 Audio-based Community Detection

Both algorithms described above are based on the adjacency matrix A of the graph. This allows us to easily extend these algorithms to include audio-based similarity measures. We simply insert an inter-node similarity value for each non-zero entry in A. We calculate these similarity values using audio-based analysis.

For the audio analysis, MFCCs are extracted resulting in 100ms non-overlapping frames. For each artist node a GMM is built from the concatenation of MFCC frames for all songs found on each artist's Myspace page (generally between 1 and 4 songs although some artists have more). For each non-zero value in the adjacency matrix A_{ij} a dissimilarity value is calculated using the earth mover's distance λ_{ij} between the GMMs corresponding to nodes i and j.

These dissimilarity values must be converted to similarity values to be successfully applied to the community detection algorithms. This is achieved by taking the reciprocal of each dissimilarity.

$$A_{ij} = \begin{cases} \lambda_{ij}^{-1} & \text{if nodes } i \text{ and } j \text{ are connected,} \\ 0 & \text{otherwise.} \end{cases} \quad (7)$$

3.3 Genre Entropy

Now that we have several methods for detecting community structures in our network, we need a means of evaluating the relevance of these structures in the context of music. Traditionally, music and music artists are classified in terms of *genre*. If the structure of the Myspace artist network is relevant to music, we would expect the communities identified within the network to be correlated with musical genres. That is, communities should contain nodes with a more homogenous set of genre associations than the network as a whole.

As mentioned in Section 3.1, we have collected genre tags that are associated with each artist. In order to measure the diversity of each community with respect to genre we use a variant of Shannon entropy we call *genre entropy S*. This approach is similar to that of Lambiotte [13]. For a given community C_k we calculate genre entropy as:

$$S_{C_k} = - \sum_{\gamma \in C_k} P_{\gamma|C_k} \log P_{\gamma|C_k} \quad (8)$$

where $P_{\gamma|C_k}$ is the probability of finding genre tag γ in community C_k. As the diversity of genre tags in a community C_k increases, the genre entropy S_{C_k} increases. As the genre tags become more homogenous, the value of S_{C_k} decreases. If community C_k is described entirely by one genre tag then $S_{C_k} = 0$. We can calculate an overall genre entropy S_G by including the entire network sample. In this way, we can evaluate each community identified by comparing S_{C_k} to S_G. If the community structures in the network are related to musical genre, we would expect the communities to contain more homogenous mixtures of genre tags. That is, in general, we would expect $S_{C_k} \le S_G$. However, as community size decreases so will the genre entropy because fewer tags are available. To account for this, we create a random partitioning of the graph that results in the same number of communities and calculate the corresponding genre entropies S_{rand} to provide a baseline.

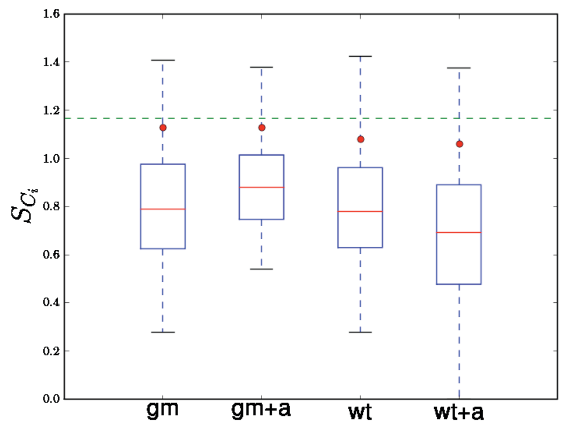

Figure 1. Box and whisker plot showing the spread of community genre entropies for each graph partition method where gm is greedy modularity, gm+a is greedy modularity with audio weights, wt is walktrap, and wt+a is walktrap with audio weights. The horizontal line represents the genre entropy of the entire sample. The circles represent the average value of genre entropy for a random partition of the network into an equivalent number of communities.

If an artist specified no genre tags, this node is ignored and makes no contribution to the genre entropy score. In our data set, 2.6% of artists specified no genre tags.

4 RESULTS

The results of the various community detection algorithms are summarized in Figure 1 and Table 1. When the genre entropies are averaged across all the detected communities, we see that for every community detection method the average genre entropy is lower than S_G as well as lower than the average genre entropy for a random partition of the graph into an equal number of communities. This is strong evidence that the community structure of the network *is* related to musical genre.

It should be noted that even a very simple examination of the genre distributions for the entire network sample suggests a network structure that is closely related to musical genre. Of all the genre associations collected for our data set, 50.3% of the tags were either "Hip-Hop" or "Rap" while 11.4% of tags were "R&B". Smaller informal network samples, independent of our main data set, were also dominated by a handful of similar genre tags (i.e. "Alternative", "Indie", "Punk"). In context, this suggests our sample was essentially "stuck" in a community of Myspace artists associated with these particular genre inclinations. However, it is possible that these genre distributions are indicative of the entire Myspace artist network. Regardless, given that the genre entropy of our entire set is so low to begin with it is an encouraging result that we could efficiently identify communities of artists with even lower genre entropies.

algorithm	c	$\langle S_C \rangle$	$\langle S_{rand} \rangle$	Q
none	1	1.16	-	-
gm	42	0.81	1.13	0.61
gm+a	33	0.90	1.13	0.64
wt	195	0.80	1.08	0.61
wt+a	271	0.70	1.06	0.62

Table 1. Results of the community detection algorithms where c is the number of communities detected, $\langle S_C \rangle$ is the average genre entropy for all communities, $\langle S_{rand} \rangle$ is the average genre entropy for a random partition of the network into an equal number of communities, and Q is the modularity for the given partition.

Without audio-based similarity weighting, the greedy modularity algorithm (gm) and the walktrap algorithm (wt) result in genre entropy distributions with no statistically significant differences. However the walktrap algorithm results in almost five times as many communities which we would expect to result in a lower genre entropies because of smaller community size. Also note that the optimized greedy modularity algorithm is considerably faster than the walktrap algorithm - $\mathcal{O}(m \log n)$ versus $\mathcal{O}(n^2 \log n)$.

With audio-based similarity weighting, we see mixed results. Applying audio weights to the greedy modularity algorithm (fg+a) actually increased genre entropies but the differences between fg and fg+a genre entropy distributions are not statistically significant. Audio-based weighting applied to the walktrap algorithm (wt+a) results in a statistically significant decrease in genre entropies compared to the un-weighted walktrap algorithm ($p = 4.2 \times 10^{-4}$). It should be noted that our approach to audio-based similarity results in dissimilarity measures that are mostly orthogonal to network structure [8]. Future work will include the application of different approaches to audio-based similarity.

5 MYSPACE AND THE SEMANTIC WEB

Since our results indicate that the Myspace artist network is of interest in the context of music-related studies, we have made an effort to convert this data to a more structured format. We have created a Web service [6] that describes any Myspace page in a machine-readable Semantic Web format. Using FOAF [7] and the Music Ontology [8], the service describes a Myspace page in XML RDF. This will allow future applications to easily make use of Myspace network

[6] available at `http://dbtune.org/myspace`
[7] `http://www.foaf-project.org/`
[8] `http://musicontology.com/`

data (i.e. for music recommendation).

6 CONCLUSIONS

We have presented an analysis of the community structures found in a sample of the Myspace artist network and shown that these community structures are related to musical genre. We have applied two efficient algorithms to the task of partitioning the Myspace artist network sample into communities and we have shown how to include audio-based similarity measures in the community detection process. We have evaluated our results in terms of genre entropy - a measure of genre tag distributions - and shown the community structures in the Myspace artist network are related to musical genre.

In future work we plan to examine community detection methods that operate locally, without knowledge of the entire network. We also plan to address directed artist graph analysis, bipartite networks of artists and listeners, different audio analysis methods, and the application of these methods to music recommendation.

7 ACKNOWLEDGEMENTS

Thanks to Zhijia Huang, Chris Sutton, Matthew Davies, Amelie Anglade and Yves Raimond for their input. This work is supported as a part of the OMRAS2 project, EPSRC numbers EP/E02274X/1 and EP/E017614/1.

8 REFERENCES

[1] Y. Ahn, S. Han, H. Kwak, S. Moon, and H. Jeong. Analysis of topological characteristics of huge online social networking services. In *WWW '07: Proceedings of the 16th international conference on World Wide Web*, pages 835–844, New York, NY, USA, 2007. ACM.

[2] A. Anglade, M. Tiemann, and F. Vignoli. Virtual communities for creating shared music channels. In *Proc. of ISMIR*, pages 95–100, 2007.

[3] A. Berenzweig, B. Logan, D. P. W. Ellis, and B. P. W. Whitman. A large-scale evaluation of acoustic and subjective music-similarity measures. *Computer Music J.*, 28(2):63–76, 2004.

[4] P. Cano, O. Celma, M. Koppenberger, and J. M. Buldu. The topology of music recommendation networks. arXiv.org:physics/0512266, 2005.

[5] O. Celma and P. Lamere. Music recommendation. ISMIR Tutorial, 2007.

[6] A. Clauset, M. E. J. Newman, and C. Moore. Finding community structure in very large networks. *Phys. Rev. E*, 70(6):066111, Dec 2004.

[7] L. F. Costa, F. A. Rodrigues, G. Travieso, and P. R. Villas Boas. Characterization of complex networks: A survey of measurements. *Advances In Physics*, 56:167, 2007.

[8] 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 *Proc. of ICMC*, August 2008.

[9] P. Gleiser and L. Danon. Community structure in jazz. *Advances in Complex Systems*, 6:565, 2003.

[10] K. Jacobson. A multifaceted approach to music similarity. In *Proc. of ISMIR*, 2006.

[11] K. Jacobson and M. Sandler. Musically meaningful or just noise, an analysis of on-line artist networks. In *Proc. of CMMR*, pages 306–314, 2008.

[12] H. Kwak, S. Han, Y. Ahn, S. Moon, and H. Jeong. Impact of snowball sampling ratios on network characteristics estimation: A case study of cyworld. Technical Report CS/TR-2006-262, KAIST, November 2006.

[13] R. Lambiotte and M. Ausloos. On the genre-fication of music: a percolation approach (long version). *The European Physical Journal B*, 50:183, 2006.

[14] S. H. Lee, P. J. Kim, and H. Jeong. Statistical properties of sampled networks. *Physical Review E*, 73:102–109, January 2006.

[15] B. Logan. Mel frequency cepstral coefficients for music modeling. In *Proc. of ISMIR*, 2000.

[16] B. Logan and A. Salomon. A music similarity function based on signal analysis. *Multimedia and Expo ICME*, pages 745–748, 2001.

[17] M. E. J. Newman. The structure and function of complex networks. *SIAM Review*, 45:167, 2003.

[18] E. Pampalk. *Computational Models of Music Similarity and their Application in Music Information Retrival*. PhD thesis, Technischen Universität Wien, May 2006.

[19] J. Park, O. Celma, M. Koppenberger, P. Cano, and J. M. Buldu. The social network of contemporary popular musicians. arXiv:physics/0609229, 2006.

[20] P. Pons and M. Latapy. Computing communities in large networks using random walks (long version). arXiv:physics/0512106v1, 2005.

[21] Y. Rubner, C. Tomasi, and L. J. Guibas. The earth mover's distance as a metric for image retrieval. *International Journal of Computer Vision*, 40(2):99–121, 2000.