SOCIAL PLAYLISTS AND BOTTLENECK MEASUREMENTS : EXPLOITING MUSICIAN SOCIAL GRAPHS USING CONTENT-BASED DISSIMILARITY AND PAIRWISE MAXIMUM FLOW VALUES

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

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

ABSTRACT

We have sampled the artist social network of Myspace and to it applied the pairwise relational connectivity measure Minimum cut/Maximum flow. These values are then compared to a pairwise acoustic Earth Mover's Distance measure and the relationship is discussed. Further, a means of constructing playlists using the maximum flow value to exploit both the social and acoustic distances is realized.

1 INTRODUCTION

As freely–available audio content continues to become more accessible, listeners require more sophisticated tools to aid them in the discovery and organization of new music that they find enjoyable. This need, along with the recent advent of Internet based social networks and the steady progress of signal based Music Information Retrieval have created an opportunity to exploit both social relationships and acoustic similarity in recommender systems.

Motivated by this, we examine the Myspace artist network. Though there are a number of music oriented social networking websites, Myspace[1] has become the *de facto* standard for web-based music artist promotion. Although exact figures are not made public, recent estimates suggest there are well over 7 million artist pages[2] on Myspace. For the purpose of this paper, *artist* and *artist page* are used interchangeably to refer to the collection of media and social relationships found at a specific Myspace page residing in Myspace's artist subnetwork, where this subnetwork is defined as those Myspace user pages containing the audio player application.

The Myspace social network, like most social networks, is based upon relational links between *friends* designating some kind of association. Further, a Myspace user has a subset of between 8 and 40 *top friends*. While all friends

are mutually confirmed, individual users unilaterally select top friends. Additionally, pages by *artists* will usually contain streaming and downloadable media of some kind either audio, video or both.

Social networks of this sort present a way for nearly anyone to distribute their own media and as a direct result, there is an ever larger amount of available music from an ever increasing array of artists. Given this environment of content, how can we best use all of the available information to discover new music? Can both social metadata and content based comparisons be exploited to improve navigation?

To work towards answers to these and related questions, we explore the relationship between the connectivity of pairs of artists on the Myspace top friends network and a measure of acoustic dissimilarity of these artists.

We begin this paper by briefly reviewing graph theoretic network flow analysis and previous work in related topics including musician networks, content-based artist similarity. We go on to explain our methodology including our network sampling method in Section 3.1 and our connectivity analysis techniques in Section 3.2. These connectivity measures are then compared to acoustic artist similarity for the structured network in Section 4 and they are used to construct a social playlist in Section 5. We finish with a discussion of the results and what these results may mean for future work in this space.

2 BACKGROUND

This work uses a combination of complex network theory, network flow analysis and signal-based music analysis. Both disciplines apply intuitively to Music Information Retrieval; however, the two have only recently been applied simultaneously to a single data set [9].

2.1 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. In contrast to *simple networks*, all these networks exhibit several unifying characteristics such as small worldness, scale-free degree distributions, and community structure [19]. We briefly introduce below some definitions and concepts that will be used in this work.

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 (i, j) it connects. This pair of nodes are *neighbors* via edge $E(i, j)$. If the edges imply directionality, $(i, j) \neq (j, i)$, the network is a *directed network*. Otherwise, it is an *undirected network*. In this paper, all edges are directed unless otherwise stated. In some graphs each edge $E(i, j)$ will have an associated label $w(i, j)$ called the *weight*. This weight is sometimes thought of as the cost of traversing an edge, or an edge's resistance. The number of edges incident to 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.

The *degree distribution $P(k)$* of a graph 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 $P(k) \propto k^{-\gamma}$, while *random* networks have a Poisson distribution [19]. Many real-world networks are approximately scale-free over a wide range of scales. Conceptually, a scale-free distribution indicates the presence of a few very-popular *hubs* that tend to attract more links as the network evolves [19].

2.2 Network Flow Analysis

The basic premise in network flow analysis is to examine a network's nodes as sources and sinks of some kind of *traffic*[2]. Typically, though not exclusively, flow networks are directed, weighted graphs. A simple flow network can be seen in Figure 1. Many useful strategies for determining the density of edge connectivity between sources and sinks can be found in this space[18]. One of the most common among them is the Maximum Flow/Minimum Cut Theorem[8], which is a means of measuring the maximum capacity for fluid to flow between a source node to a sink node or, equivalently, the smallest sum of edge weights that must be *cut* from the network to create exactly two subgraphs, one containing the source node and one containing the sink node. This will be discussed in more detail in Section 3.2. The few examples of network flow type analysis in music deal primarily with constructing playlists using partial solutions to the Traveling Salesman Problem [12] or use exhaustive and explicit metadata[3].
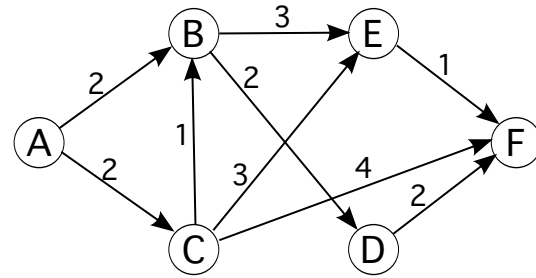
Figure 1. A simple flow network with directed weighted edges. Here the source is node A and the sink is node F.

2.3 Musician Networks

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 [7, 10, 21]. However, these studies all examine networks created by experts (*e.g.* All Music Guide [3]) or via algorithmic means (*e.g.* Last.fm [4]) as opposed to the artists themselves, as is seen in Myspace and other similar networks. Networks of music listeners and bipartite networks of listeners and artists have also been studied [4, 14].

2.4 Content-Based Music Analysis

Many methods have been explored for content-based music analysis, attempting to 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 [16]. In combination with various statistical techniques, MFCCs have been successfully applied to music similarity and genre classification tasks [6, 17, 20]. A common approach for computing timbre-based similarity between two songs or collections of songs creates Gaussian Mixture 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 [22], is the distance measure used for this purpose [5, 20]. The EMD algorithm finds the minimum work required to transform one distribution into another.

2.5 Bringing It Together

There has been some recent work attempting to bridge the divide between content-based analysis and human generated metadata. Most of this work [12, 23] focuses on various ways of exploiting the human-generated metadata to filter content prior to, or instead of, conducting content-based analysis, similar to the techniques discussed in Section 2.4, in order to reduce computational load.

3 METHODOLOGY

3.1 Sampling the Social Web

The Myspace social network presents a variety of challenges. For one, the size of the network prohibits analyzing the graph in its entirety, even when considering only the artist pages. Therefore in the work we deal with a sample (of large absolute size) of the network. 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 [5]. Our data is collected using web crawling and HTML scraping techniques [6].

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 most often 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 sampled network but also artificially limits the outdegree of each node. Our sampling is based on the assumption that artists specified as top friends have some meaningful musi-

cal connection for the user – whether through collaboration, stylistic similarity, friendship, or artistic influence.

The audio files associated with each artist page in the sampled network are also collected for feature extraction. Cached versions of the audio files are downloaded and audio features are extracted.

3.1.2 Snowball Sampling

There are several network sampling methods; however, for the Myspace artist network, snowball sampling is the most appropriate method [1, 15]. In this method, the sample begins with a seed node (artist page), then the seed node's neighbors (top friends), then the neighbors' neighbors, are added to the sample. This breadth-first sampling is continued until a particular sampling ratio is achieved. Here, we randomly select a seed artist [7] and collect all artist nodes within 6 edges to collect 15,478 nodes. This produces a dataset where no more than six directed top friends links need to be followed to get from the seed artist to any other artist in the dataset. If the size of the Myspace artist network is around 7 million, then this dataset approximates the 0.25% sampling ratio suggested for accurate degree distribution estimation in sampled networks. However, it is insufficient for estimating other topological metrics such as the clustering coefficient and assortativity [13]. Of course, a complete network topology is not our primary concern here.

With snowball sampling there is a tendency to over sample hubs because they have high indegree connectivity and are therefore picked up disproportionately frequently in the breadth-first sampling. This property would reduce the degree distribution exponent γ and produce a heavier tail but preserve the power-law nature of the network [15].

3.2 Minimum Cut/Maximum Flow

We use the Maximum Flow value as a means of determining the number of independent paths from a source node to a sink node. Formally the Maximum Flow/Minimum Cut theorem[8], it is used to calculate the highest weight in the narrowest part of the path from source to sink. The theorem's name comes from the equivalence in the smallest weight of edges that must be removed in order to create two subgraphs which disconnect the source and sink nodes. Further, if the edges in the graph are unweighted, this value is also equivalent to the number of paths from the source to the sink which share no common edges. As this is a mature algorithm there are a number of optimization strategies that have been examined [2, 11].

An example of Maximum Flow can be seen on the network in figure 1. It can been seen that the narrowest point

---

[5] At time of writing Myspace has recently published a public API, this may allow future work to occur without the need for html scraping, which would greatly decrease the compute time required for graph generation.

[6] Myspace scraping is done using tools from the MyPySpace project available at `http://mypyspace.sorceforge.net`

[7] The artist is *Karna Zoo*, Myspace url: `http://www.myspace.com/index.cfm?fuseaction=user.viewProfile&friendID=134901208`

from node A to node F is $E(a, b)$ and $E(a, c)$. The maximum flow can simply be found via Equation 1.

$$M = \sum m(i, j) \qquad (1)$$

Where $m(i, j)$ is the magnitude of each edge in the minimum cut set.

In our Myspace top friends graph, the maximum flow is measured on the unweighted directed graph from the source artist node to the sink artist node.

4 CONNECTED SOUND

4.1 Experiment

We calculate the maximum flow value, using the snowball sample entry point as the fixed source against every other node in turn as a sink, yielding the number of edges connecting each sink node to the entry point node at the narrowest point of connection. The acoustic distances can then be compared to these maximum flow values.

4.1.1 Signal-based analysis

Cepstral coefficients are extracted from each audio signal using a Hamming window on 8192 sample FFT windows with 4096 sample overlap. For each artist node a GMM is built from the concatenation of MFCC frames for all songs found on each artist's Myspace page (the mean number of songs per artist is 3.5). We calculate the Earth Mover's Distance between the GMMs corresponding to each source sink pair in the sample. All MFCCs are created with the fftExtract tool [8].

4.1.2 Random Networks

In order to better understand a result from analysis of our Myspace sample, a baseline for comparison must be used. To that end, random permutations of the node locations are examined. In order to preserve the overall topology present in the network, this randomization is performed by randomizing the artist label and associated music attached to a given node on the network. This is done ten fold, creating a solid baseline to test the null hypothesis that the underlining community structure is not responsible for any correlation between maximum flow values and Earth Mover's Distance.

4.2 Result

The results of the first experiment show no simple relationship between the sampled netwrok and the randomized network. This can be seen in Table 1 and in Figures 2 and 3. There is an increase in the median EMD for the less well connected (*i.e.* lower maximum flow value) node pairs in the

---

[8] source code at *http://omras2.doc.gold.ac.uk/software/fftextract/*

Max Flow	median	deviation	randomized	deviation
1	40.80	1.26	39.10	−0.43
2	45.30	5.76	38.34	−1.19
3	38.18	−1.35	38.87	−0.66
4	38.21	−1.32	38.64	−0.89
5	40.00	0.47	39.11	−0.42
6	41.77	2.25	39.02	−0.51
7	39.94	0.41	39.24	−0.29
8	39.38	−0.15	38.76	−0.77
9	38.50	−1.03	38.87	−0.66
10	39.07	−0.46	40.85	1.32

Table 1. Node pairs average EMD values grouped by actual minimum cut values and randomized minimum cut values, shown with deviations from the global median of 39.53.

Myspace sample graph, though this is not significant enough to indicate a correlation, while the randomized permutations are near flat. While the median EMD of the artist pairs with a maximum flow of 10 is appreciably higher than all other in the randomized graph, this is likely related to the relatively large size of this group. Perhaps the easiest way to examine the relationship between the sampled graph and randomized one is through the deltas of each group's median from the entire dataset median. This data is shown in the second and forth colum in Table 1 and Figure 4. Further, the Kruskal-Wallis one-way ANOVA results for both the sample graph and averaged across the 10 fold permutations are shown in Table 2.

	H-value	P-value
From sample	12.46	0.19
Random permutations	9.11	0.43

Table 2. The Kruskal-Wallis one-way ANOVA test results of EMD against maximum flow for both the sampled graph and it's random permutations. The H-values are drawn from a chi-square distribution with 10 degrees of freedom.

5 THE MAX FLOW PLAYLIST

In order to build playlists using both acoustic and social network data, we use the Earth Mover's Distance between each pair of neighbors as weights on the Myspace sample network. Two artists are then selected, a starting artist as the source node and a final artist as the sink node. One or more paths are then found through the graph via the maximum flow value, generating the list and order of artists for the playlist. The song used is the most popular at the time of the page scrape. In this way playlists are constructed that are both influenced by timbre similarity and bound by so-
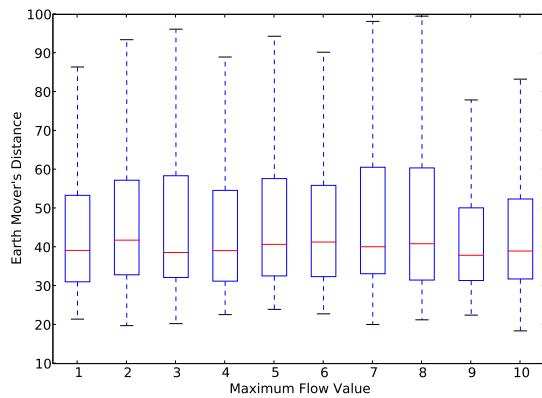
Figure 2. The box and whisker plot showing the distribution of EMD grouped by maximum flow value between artists on the Myspace social graph.
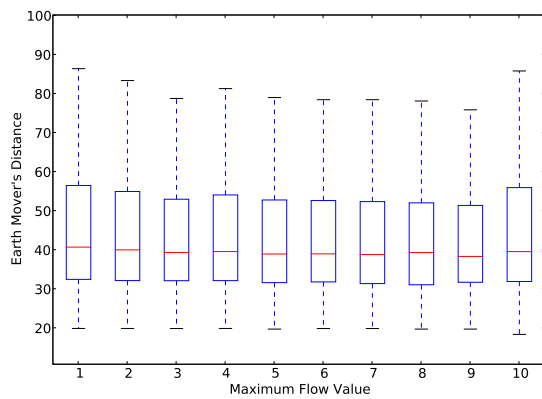

Figure 3. The box and whisker plot showing the distribution of EMD grouped by maximum flow value between artists on the randomized graph, maintaining the original edge structure.

cial context, regardless of any relationship found between these two spaces found via the work discussed in Section 4. Playlists generated using this technique were informally auditioned, and were found to be reasonable on that basis.

6 DISCUSSION AND FUTURE WORK

While an inverse relationship between Earth Mover's Distance and the maximum flow value might be expected on the basis of the conventional wisdom that a community of artists tend to be somehow aurally similar, this does not appear to be strictly the case. The evidence, at least in this sample set, does not support this relationship, though it doesn't disprove it either. However, based upon the difference in result from the Kruskal-Wallis one-way ANOVA test and simple obser-
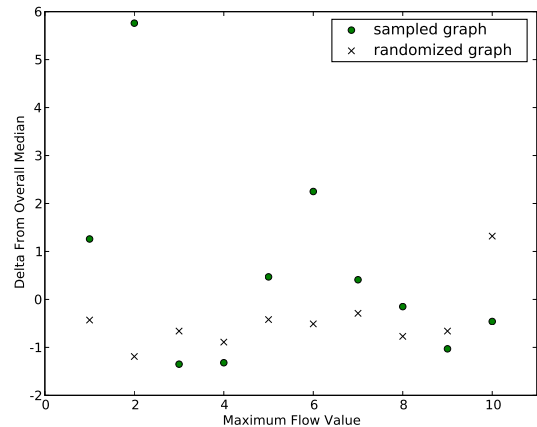
Figure 4. The deltas from the global median for each maximum flow value group of EMD values, from the sampled graph and the randomized graph.

vation of the deviation from the global median the maximum flow values and Earth Mover's Distances do seem affected by the artist created social links, though it is not a simple relationship and describing it precisely is difficult. This seems to suggest that there may be no noticeable correlation between density of connection (maximum flow values) and acoustic similarity (GMMs compared with EMD), at least across an entire sample.

There does seem to be some potential in the idea of the maximum flow playlist. When using the EMD as a weight the results appear to be quite good, at least from a qualitative perspective. The imposed constraint of the social network alleviates to some extent short comings of a playlist built purely through the analysis of acoustic similarity by moving more toward the balance between completely similar works and completely random movement.

Having shown the lack of a strong relationship between the maximum flow values and acoustic artist similarity, where do we go from here?

The most promise lies in the exploration of the maximum flow based playlist. A network could be built which was song to song exhaustive, having duplicate edges link each song individually to an artist's friends' songs. These edges would be weighted according to their acoustic similarity and a more complete playlist generation system would be created. A serious hurdle to the implementation of such a system lies in the computational complexity of the maximum flow value. Its compute time is typically dependent on both the number of nodes and the number of edges making it very slow to run on a network as dense as the one just described. This is less of a concern if some form of localized subgraphs were used, *e.g.* maximum flow is found against only friends (the *greedy* approach) or friends of friends. That said, there

563

may be strategies to get around these problems of complexity leading to novel and interesting playlist generation.

7 ACKNOWLEDGEMENTS

8 REFERENCES

[1] 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, May 2007, IW3C2.

[2] R. K. AHUJA, T. L. MAGNANTI, AND J. B. ORLIN, *Network flows: theory, algorithms, and applications*, Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 1993.

[3] M. ALGHONIEMY AND A. TEWFIK, *A network flow model for playlist generation*, in Multimedia and Expo, 2001. IEEE International Conference on, 2001.

[4] A. ANGLADE, M. TIEMANN, AND F. VIGNOLI, *Virtual communities for creating shared music channels*, in Proc. of Int. Symposium on Music Information Retrieval, 2007.

[5] J. AUCOUTURIER AND F. PACHET, *Improving timbre similarity : How high's the sky?*, J. Negative Results in Speech and Audio Sciences, (2004).

[6] 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 (2004), pp. 63–76.

[7] P. CANO, O. CELMA, M. KOPPENBERGER, AND J. M. BULDU, *The topology of music recommendation networks*, Chaos: An Interdisciplinary Journal of Nonlinear Science, (2006). http://arxiv.org/abs/physics/0512266v1.

[8] P. ELIAS, A. FEINSTEIN, AND C. SHANNON, *A note on the maximum flow through a network*, Information Theory, IEEE Transactions on, 2 (Dec 1956), pp. 117–119.

[9] 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 Int. Computer Music Conference, August 2008.

[10] P. GLEISER AND L. DANON, *Community structure in jazz*, Advances in Complex Systems, 6 (2003), pp. 565–573.

[11] A. V. GOLDBERG AND R. E. TARJAN, *A new approach to the maximum-flow problem*, J. ACM, 35 (1988), pp. 921–940.

[12] P. KNEES, T. POHLE, M. SCHEDL, AND G. WIDMER, *Combining audio-based similarity with web-based data to accelerate automatic music playlist generation*, in Proc. 8th ACM international workshop on Multimedia information retrieval, 2006, pp. 147 – 154.

[13] 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*, Tech. Rep. CS/TR-2006-262, KAIST, November 2006.

[14] R. LAMBIOTTE AND M. AUSLOOS, *On the genrefication of music: a percolation approach (long version)*, The European Physical Journal B, 50 (2006), p. 183.

[15] S. H. LEE, P.-J. KIM, AND H. JEONG, *Statistical properties of sampled networks*, Physical Review E, 73 (2006), pp. 102–109.

[16] B. LOGAN, *Mel frequency cepstral coefficients for music modeling*, in Proc. of Int. Symposium on Music Information Retrieval, 2000.

[17] B. LOGAN AND A. SALOMON, *A music similarity function based on signal analysis*, in Multimedia and Expo, 2001. IEEE International Conference on, 22-25 Aug. 2001, pp. 745–748.

[18] H. NAGAMOCHI AND T. IBARAKI, *Computing edge-connectivity in multigraphs and capacitated graphs*, SIAM J. Discret. Math., 5 (1992), pp. 54–66.

[19] M. E. J. NEWMAN, *The structure and function of complex networks*, SIAM Review, 45 (2003), p. 167.

[20] E. PAMPALK, *Computational Models of Music Similarity and their Application in Music Information Retrieval*, PhD thesis, Technischen Universität Wien, May 2006.

[21] J. PARK, O. CELMA, M. KOPPENBERGER, P. CANO, AND J. M. BULDU, *The social network of contemporary popular musicians*, Int. J. of Bifurcation and Chaos, 17 (2007), pp. 2281–2288.

[22] 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 (2000), pp. 99–121.

[23] M. SLANEY AND W. WHITE, *Similarity based on rating data*, in Proc. of Int. Symposium on Music Information Retrieval, 2007.