LISTEN TO ME – DON'T LISTEN TO ME: WHAT CAN COMMUNITIES OF CRITICS TELL US ABOUT MUSIC

Ben Fields, Christophe Rhodes
Dept. of Computing
Goldsmiths University of London
New Cross
London SE14 6NW
United Kingdom
me@benfields.net, c.rhodes@gold.ac.uk

ABSTRACT

Social knowledge and data sharing on the Web takes many forms. So too do the ways people share ideas and opinions. In this paper we examine one such emerging form: the amateur critic. In particular, we examine genius.com, a website which allows its users to annotate and explain the meaning of segments of lyrics in music and other written works. We describe a novel dataset of approximately 700,000 users' activity on genius.com, their social connections, and song annotation activity. The dataset encompasses over 120,000 songs, with more than 3 million unique annotations. Using this dataset, we model overlap in interest or expertise through the proxy of co-annotation. This is the basis for a complex network model of the activity on genius.com, which is then used for community detection. We introduce a new measure of network community activity: community skew. Through this analysis we draw a comparison of between co-annotation and notions of genre and categorisation in music. We show a new view on the social constructs of genre in music.

1. INTRODUCTION

The near-ubiquitous availability and use of the Web has enabled many otherwise dispersed communities to coalesce. Many of these communities are concerned with the gathering and transfer of knowledge. Perhaps the best known of this kind of community is that of the editors and contributors at Wikipedia [1] [16, 22]. However, people coming together in a shared virtual space to exchange ideas is not limited to curation of encyclopedic facts. The Web is full of many communities; this paper focuses on an emerging one with a particular relevance to music: genius.com [2].

Genius.com brings users together through *annotation*. The stated purpose, and indeed, general use of the site is to explain portions of text through annotating them. These annotations can themselves be edited and modified, much as would take place on a website such as Wikipedia. Unlike on Wikipedia, however, the goal of allowing annotations is specifically to generate metadata: These annotations are both opinion and derivative works, criticism for the twitter age.

We have collected a significant sample of the user activity on Genius. This sample forms the core of a dataset that is ripe with potential. To show this, we construct a bipartite graph model of our Genius sample, connecting users and works via annotations made on those works. This graph model is then used to compare the communities formed around annotation with the genre prescribed to the annotated works. In doing this we seek to test the fitness and cultural relevance of the prescribed genre to these works.

The remainder of this paper is organized into the following sections. In Section 2 we discuss the relevant contexts: social network analytics in general, specific work in music, complex networks and community detection. From there, in Section 3 we describe the dataset – the collection techniques along with various statistics concerning the raw captured data. In Section 4 we then explore one possible avenue of use of our dataset, network modelling and community detection. We look at how detected communities align with prescribed genre labels for the works in these communities with a novel metric, *community skew*. Finally, we state our conclusions and consider what the next steps should be in Section 5.

2. BACKGROUND

When considering a social network of criticism such as Genius, we must consider what the landscape looks like to place this work in a more complete context.

[1] http://wikipedia.org

[2] The website and company began as rapgenius (http://rapgenius.com) with a strong focus on explaining the nuance, references, and in-jokes of rap and hip-hop lyrics. However they rebranded as 'Genius' as they widened their focus, which now includes lyrics from all genre of music as well as poetry, libretti, and factual texts such as news articles. See this announcement from 12 July 2014 http://genius.com/Genius-founders-introducing-geniuscom-annotated.

 Attribution: Ben Fields, Christophe Rhodes. "Listen to Me – Don't Listen to Me: What Can Communities of Critics Tell Us About Music", 17th International Society for Music Information Retrieval Conference, 2016.

2.1 Music and Social Networks

While Genius has existed in some capacity since July of 2010 [3], it is one of many social networks with user-generated content (UGC) and an emphasis on music. One of the earliest of these networks was youtube [4]. While youtube is ostensibly a site for hosting and sharing video, it is also the single most prolific source of music on the Web [5]. Further, its social structure was one of the first on the modern Web to be extensively studied [4, 17]. It was shown that youtube, like many other Web-based social networks, has a power-law roll off in the distribution of its users' connections to other users and that the users congregate into clumps of tightly connected communities, showing 'small-world' characteristics.

Other Web-based communities brought together content creators with a greater explicit emphasis on social connections. In particular, myspace [6] has been looked at, both in terms of community structures [13] and as a proxy for understanding song and artist similarity [6, 7]. Further, these techniques have been used to drive recommenders and playlist generation [8]. In recent years, Soundcloud [7] has become the Web platform of choice for this combination of audio recordings and social network connectivity. It has broadly similar network characteristics [12, Chapter 3] with its own particular traits, reflecting interface and design decisions as well as the different user composition of the network. In addition to these networks around complete works, analysis has been done showing associations between properties of the contributor network for Freesound [8] (an open collection of audio clips) and creative outcomes among participants [19].

Analogous work has also been done on the listener or consumer side. In particular various aspects of listening and sharing behaviour on twitter [9] have been studied. The twitter microblogging platform has been successfully used to model artist similarity and descriptors, based on network ties and other attributes [20]. Extensions of this work then used twitter to show popularity trends across both time and space [21]. Going a step further, twitter network analysis can be used to create and order personalized playlists [14].

2.2 Information and Social Networks

While a significant volume of research has been done on information gathering social networks, it nearly exclusively uses Wikipedia as the source social network. As mentioned in Section 1, Wikipedia aims to be encyclopedic in both tone and scope, which colours the network significantly. Nevertheless, this work can offer useful insight and approaches for networks of this type.

Complex network techniques are effective in determining the most influential nodes across an information network [15]. This can be used to help understand how information flows through a social network. Wikipedia editors can broken down into different classes based on their behaviour within the network [11].

3. THE DATASET

In this section we describe the general structure of Genius, especially as it pertains to the dataset presented in this paper. We go into detail about the process of scraping and spidering the site to collect the data, highlighting sampling decisions and noting possible biases. Lastly, we present a statistical overview of the features of the dataset.

3.1 The Structure of Genius

At its core Genius is a collection of textual representations of works, most commonly but not exclusively lyrics. Each of these works are rendered such that an arbitrary sequence of words may be selected and a user may then write some commentary about the meaning of this section of the work (the *annotation*). An example of this display can be seen in Figure 1, in this case lyrics for *Hypnotize* by The Notorious B.I.G. with the line 'Timbs for hooligans in Brooklyn' highlighted with the annotation visible.

Once an annotation has been placed by a user, it can be edited and debated. This process can involve significant back and forth between users, as those interested within community voice their point of view as to the meaning of a line. The result is an annotation that reflects a collective process: the contributions that have led to the current state of an annotation are easily viewable, as can be seen in Figure 2 with the same annotation as the previous figure.

A user maintains a profile on Genius, as is the case on many social networks. Central to this profile is the history of the annotations made by the user. As such, a user's persona on Genius is effectively the collection of their annotations across the site. One such user profile is shown in Figure 3, that of the user 'OldJeezy', the lead contributor to the previously mentioned on annotation for the work *Hypnotize*.

3.2 Collecting the Data

Until recently Genius lacked any kind of machine-readable API [10], so our data collection effort restructured data drawn from the html as presented to a user. The data collection efforts on Genius are made up of two parts: a spider and a scraper. The spider, or mechanism to automatically move through the pages to be collected, sets out to evenly sample across the space of user IDs, without preference for or against how active a particular users is on the site. This algorithm is reasonably straight-forward and relies on the

[3] The beginning of their current site can be seen dating back to 22 July 2010 according to `http://web.archive.org/web/20100615000000*/http://rapgenius.com`

[4] `http://youtube.com`

[5] `http://www.nielsen.com/us/en/press-room/2012/music-discovery-still-dominated-by-radio--says-nielsen-music-360.html`

[6] `http://myspace.com`, though it has decayed a great deal from its peak of activity circa 2006-2008

[7] `http://soundcloud.com`

[8] urlhttp://www.freesound.org/

[9] `http://twitter.com`

[10] The recently announced API (`https://docs.genius.com/`) mitigates the most of the need for further scraping via html, though the spidering and sampling techniques detailed here are unchanged.

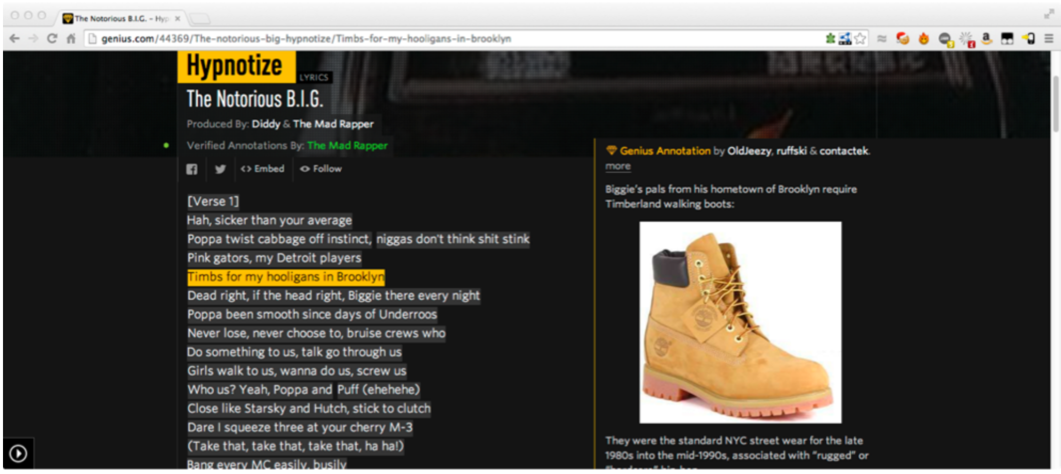

Figure 1. The lyrics to Hypnotize by The Notorious B.I.G., with an annotation shown for the line 'Timbs for hooligans in Brooklyn'. Taken from `http://genius.com/44369/The-notorious-big-hypnotize/Timbs-for-my-hooligans-in-brooklyn` on 10 March 2015.

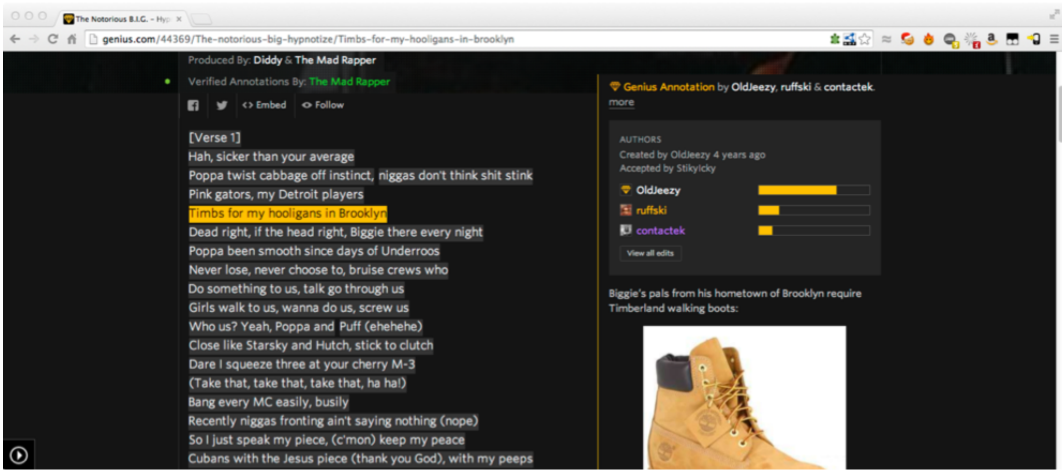

Figure 2. The same lyrics annotation as in Figure 1, but showing the total contribution of the three users who have edited the annotation for the highlighted text.

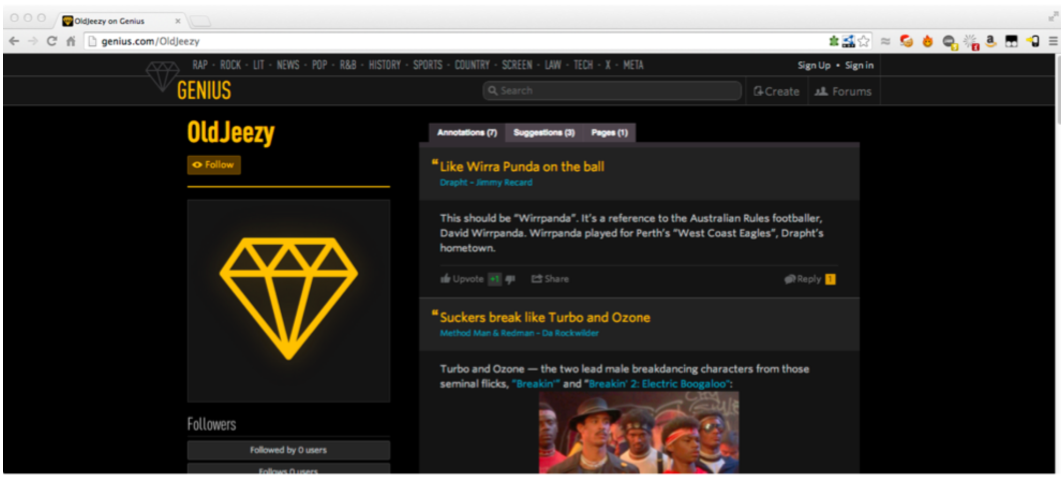

Figure 3. The recent annotation history for the user 'OldJeezy', the top contributor to the lyrics annotation shown in Figure 1. Taken from `http://genius.com/OldJeezy` on 10 March 2015

fact that Genius has sequential integer user IDs. As these ID are very nearly continuous from 0 to the most recent ID assigned, it is trivial to approximate a fair draw random generator to visit the user annotation history pages. Because of the flat and random mechanism in this spider, a partial sample is far less likely to introduce a bias toward a more densely connected graph than spidering methods that move from one user to another via a common edge (in this case a mutually annotated work). This implies that a partial capture will be reasonably representative of the whole userbase. The corresponding drawback is that any particular work may not have its entire annotation record collected, so its relative position in the topography of the graph (e.g. in terms of degree) may not be accurate, though this problem will decrease as more of the graph is captured.

To gather the data for each individual page, we created a screen scraper using Python and the BeautifulSoup [11] toolkit. This scraper is released with an open source license and is available from github [12] .

The spider and scraper were run during December 2014 collecting user metadata, annotation, works, and works metadata from the contributions of 704,438 users. This sample covers 41.1% of the 1,713,700 users [13] .

This dataset is available for download and reuse, as both CSV and SQL dump from the Transforming Musicology dataset repository [14] .

3.3 Statistical Overview

A variety of statistics describing the Genius data set can be seen in Table 1. As previously mentioned, the dataset covers the contributions of 704,438 users: 1,256,912 annotations on 146,186 unique works. Genius, as is common among many social networks [2], appears to have a steep drop off from users who sign up to users who do anything. This can be seen in the disparity between the total captured users and the contributing users (704,438 versus 71,129): 10.1% of users have written an annotation.

description	count
total users	704,438
total annotations	1,256,912
total works	146,186
contributing users	71,129
annotation edits	2,196,522
annotations with multiple contributors	194,795

Table 1. High-level statistics for Genius dataset.

Our dataset covers some 146,186 unique works and 1,256,912 annotations, giving a mean average of 8.6 annotations per work. Further, the dataset contains a total of 2,196,522 distinct edits of annotations, giving the mean annotation 1.75 edits, including its first.

Genius has 15 top-level categories for works on the site. Each work is assigned exactly one category, which can be taken as the work's genre. While that is not quite right for the non-musical categories, it is a helpful approximation. The breakdown of the works per category (genre) are seen in Table 2. The first thing that pops out is that while the company behind Genius may have decided to drop 'rap' from their name, it still dominates their collection of works, making up almost three-quarters of our dataset. While the meanings of most of these genre names are fairly typical, it is worth commenting on the few that are particular to Genius: 'x' is used as a catch-all or miscellaneous; 'screen' is for screenplays and teleplays; 'history' is for both scholarly and lay texts of a historical nature; 'unbranded' means our scraper was unable to capture the assigned genre; 'tech' covers prose about technology and the tech industry; finally 'meta' is where contributors to Genius discuss rules and community standards.

category	works count	percentage
rap	107270	73.3%
rock	16393	11.2%
lit	9386	6.2%
news	3720	2.5%
pop	3715	2.5%
sports	1140	0.7%
x	1014	0.6%
country	744	0.5%
screen	697	0.4%
r-b	655	0.4%
history	502	0.3%
unbranded	370	0.2%
law	250	0.1%
tech	159	0.1%
meta	151	0.1%

Table 2. Genre breakdown for Genius dataset.

In addition to the top-level categories, Genius supports work-level social tags. The tags have also been captured in our data set for all the works. As is typical for tags any number may be used per work, though the top-level genre category is repeated as a tag mechanically, so each work has at least one. Including these top-level categories, our dataset contains 802 unique tags. The top 10 tags (not including the categories), along with the count of the works they've been applied to, appears in Table 3.

4. NETWORK ANALYTICS

In an effort to understand what the community of annotators on Genius can tell us about that material they're annotating, we model our dataset as a graph. We use this graph, and a transform of it, to observe the community structure of works based on *co-annotation* and *user-overlap* patterns. Here co-annotation is when a common user annotates a pair of works. Similarly, user-overlap is when any pair of

[11] `http://www.crummy.com/software/BeautifulSoup/bs4/doc/`

[12] `http://dx.doi.org/10.5281/zenodo.17515`

[13] The total user count is an approximation based on the highest successfully resolving user ID as on 29 April 2015.

[14] Specifically `http://genius-annotations.data.t-mus.org/`, note that this dataset does not contain the source lyrics, only the network structure around the lyrics and their annontations. This is done for reasons of copyright compliance

category	works count
Rap Genius France	9135
Genius France	6009
Deutscher Rap	5725
Polski Rap	3298
West Coast	1384
Brasil	841
Bay Area	839
Indie Rock	716
Chicago	710
Genius Britian	540

Table 3. Top tags in Genius dataset.

users contribute to any annotation on the same work (not necessarily the same annotation).

4.1 The Graph Model

We initially model the dataset as a bipartite graph [9]. That is a graph where each node represents one of two distinct classes: a work or a user. The edges is this graph are formed whenever a user has contributed at least one annotation to a work. No edges join two nodes of the same class.

Given this graph we can discuss its topological features [1]. The graph has 216,943 nodes across both class – 71,129 of those nodes represent all the users that have contributed an annotation, 145,814 represent the works [15] . The graph contains 439,835 edges, representing the number of unique user-work pairs with annotations. While nearly half a million in number, this is quite sparse representing only 4.24×10^{-5} of the more than 10 billion possible pairs. Therefore the graph has a average degree of 2.02. This bipartite graph, serialised as graphml, is also included as part of the dataset and is available for download as mentioned in Section 3.

The remained of this section concerns the detection of communities of works. In order to do this, we project our bipartite graph to a songs-as-nodes single class graph with weighted edges representing the users that co-annotated linked works. We also only consider the largest connected component, i.e. the largest number of works for which there is a path between each pair of works included. This reduces the number of nodes to 125,044.

4.2 Examining Community

We have performed community detection with three different algorithms: fast greedy [5], leading eigenvector [18], and multilevel [3]. In order to assess the suitability of each of these inferred community structures, we take the *modularity* of each. Here modularity is a measure of the ratio of connections within communities against connections among communities. The optimum modularity resulting from each of these communities detection methods, along with the number of detected communities that give said modularity, can be seen in Table 4. Based on modularity, the multilevel community detection measure gives the best grouping, resulting in 169 distinct detected communities.

[15] The careful reader may notice that this is 372 works fewer than the 146,186 works reference in Table 1. This is due to those works URLs not resolving at the time of the crawl, most likely due to deletion of the work from the collection after the annotation was made.

method	modularity	communities
Fast greedy	0.529	498
Leading eigenvector	0.003	11488
Multilevel	**0.582**	169

Table 4. The optimum modularity scores of each of the three community detection methods used on the works graph. The highest modularity, achieved the multilevel method, is shown in bold.

category	community count	community skew
meta	16	90.0
law	12	70.0
tech	12	70.0
history	19	36.6
screen	24	35.5
r-b	23	35.0
x	24	23.3
country	19	22.0
sports	19	15.7
pop	38	8.8
news	35	8.4
unbranded	22	6.5
lit	59	5.6
rock	50	2.7
rap	143	1.2

Table 5. The spread of each genre, across detected communities.

While the higher modularity of the multilevel method is inline with previous research on other small-world graphs, the low score and high number of communities generated by the leading eigenvector method is notable and merits further investigation.

Given the 169 detected communities of works, we can compare these communities to the prescribed genre labels to see how (and if) they align. To do this, we generate a confusion matrix, analogous to what might be used to evaluate a automatic classification task. However, unlike in a common classification task, our confusion matrix is not square, having dimension so of 15 x 169 (the number of categories by the number of communities). Given the size of the confusion matrix, it is not practical to visualize the entire thing, rather we will consider it in the following reduction. [16] Since there are more than 10 times the detected communities as there are genre categories, we can see how widespread each genre category is across communities. That is, how many communities have more than zero works from a given genre. This can be seen in Table 5, which shows that spread seems to correspond with popularity of the genre label.

[16] The raw confusion matrix is available for download as a csv at `http://genius-annotations.data.t-mus.org/`

Beyond the raw counts, we can examine the *community skew* of a category S_c which we define as

$$S_c = \frac{W_c}{W} * \frac{C_c}{C} \quad (1)$$

where W_c is the number of works in category c, W is the total number of works in the corpus, C_c is the number communities in the split with at least one work category c amongst its members, and C is the total number of communities in across the network. Community skew therefore gives a measure of how widely distributed a given category label is across communities, normalised to how popular that label is in the corpus. A community skew of 1 means that the number of communities covering a genre exactly mirrors its overall representation in the corpus. Further as the skew increase away from 1 it show a disproportionate capture of the communities across the network. Looking at the community skew in Table 5 this is especially the case for the meta, law, and tech categories. With a few exceptions, the more well represented in the dataset a genre is the less its skew. This relationship implies that with more works in a genre community annotators become more distinct.

5. CONCLUSIONS AND FUTURE WORK

We have introduced the Web community Genius, a collection of (mostly music related) textual works and criticism in the form of annotations. We described and data gathering methodology, and using that methodology, collect the annotation and works metadata for the activity of over 700,000 users, with just over 10% of them active contributors. We then modelled this dataset as a bipartite graph of works and users. This graph was then projected into a single class for community detection. When performing community detection, the multilevel method was found to perform best, with a modularity score of 0.582 finding 169 communities. Using these communities we examined the community skew of each genre across these communities of works. In these community measures, and skew in particular, we see that a genre's definition is clearer as it is more popular.

While there are many further avenues of research to take this dataset and these foundations in the future, one in particular stands out: hybrid-methods using content. Performing content analysis on the lyrics, such as reading comprehension or rhyme structure analysis [10], and then using the result in tandem with cultural structures as captured by this work's network models present many possible further insights to the organisation of music.

6. ACKNOWLEDGEMENTS

This work is supported in part via the Transforming Musicology UK AHRC grant, under the Digital Transformations in the Arts and Humanities scheme, number AH/L006820/1. We would also like to acknowledge and thank the anonymous peer reviewers for their help in refining this paper.

7. REFERENCES

[1] Yong-Yeol Ahn, Seungyeop Han, Haewoon Kwak, Sue Moon, and Hawoong Jeong. Analysis of topological characteristics of huge online social networking services. In *The World Wide Web Conference (WWW)*, Alberta, Canada, May 2007.

[2] Fabrício Benevenuto, Tiago Rodrigues, Meeyoung Cha, and Virgílio Almeida. Characterizing user behavior in online social networks. In *Proceedings of the 9th ACM SIGCOMM conference on Internet measurement conference*, pages 49–62. ACM, 2009.

[3] Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10):P10008, 2008.

[4] Xu Cheng, Cameron Dale, and Jiangchuan Liu. Statistics and social network of youtube videos. In *Quality of Service, 2008. IWQoS 2008. 16th International Workshop on*, pages 229–238. IEEE, 2008.

[5] Aaron Clauset, M. E. J. Newman, and Cristopher Moore. Finding community structure in very large networks. *Physical Review E*, 70(6):066111, Dec 2004.

[6] Ben Fields, Kurt Jacobson, Michael Casey, and Mark Sandler. Do you sound like your friends? exploring artist similarity via artist social network relationships and audio signal processing. In *International Computer Music Conference (ICMC)*, August 2008.

[7] Ben Fields, Kurt Jacobson, Christophe Rhodes, and Michael Casey. Social playlists and bottleneck measurements : Exploiting musician social graphs using content-based dissimilarity and pairwise maximum flow values. In *International Conference on Music Information Retrieval (ISMIR)*, September 2008.

[8] Ben Fields, Kurt Jacobson, Christophe Rhodes, Mark Sandler, Mark d'Inverno, and Michael Casey. Analysis and exploitation of musician social networks for recommendation and discovery. *IEEE Transactions on Multimedia*, PP(99):1, 2011.

[9] Jean-Loup Guillaume and Matthieu Latapy. Bipartite graphs as models of complex networks. *Physica A: Statistical Mechanics and its Applications*, 371(2):795–813, 2006.

[10] Hussein Hirjee and Daniel G Brown. Automatic detection of internal and imperfect rhymes in rap lyrics. In *International Society on Music Information Retrieval Conference (ISMIR)*, pages 711–716, 2009.

[11] Takashi Iba, Keiichi Nemoto, Bernd Peters, and Peter A Gloor. Analyzing the creative editing behavior of wikipedia editors: Through dynamic social network analysis. *Procedia-Social and Behavioral Sciences*, 2(4):6441–6456, 2010.

[12] Kurt Jacobson. *Connections in Music*. PhD thesis, Centre for Digital Music, Queen Mary University of London, 2011.

[13] Kurt Jacobson, Ben Fields, and Mark Sandler. Using audio analysis and network structure to identify communities of on-line social networks of artists. In *International Conference on Music Information Retrieval (ISMIR)*, Philadelphia, PA, USA, October 2008.

[14] Sanghoon Jun, Daehoon Kim, Mina Jeon, Seungmin Rho, and Eenjun Hwang. Social mix: automatic music recommendation and mixing scheme based on social network analysis. *The Journal of Supercomputing*, pages 1–22, 2014.

[15] Masahiro Kimura, Kazumi Saito, and Ryohei Nakano. Extracting influential nodes for information diffusion on a social network. In *AAAI*, volume 7, pages 1371–1376, 2007.

[16] Andrea Lancichinetti, Mikko Kivelä, Jari Saramäki, and Santo Fortunato. Characterizing the community structure of complex networks. *PloS one*, 5(8):e11976, 2010.

[17] Alan Mislove, Massimiliano Marcon, Krishna P Gummadi, Peter Druschel, and Bobby Bhattacharjee. Measurement and analysis of online social networks. In *Proceedings of the 7th ACM SIGCOMM conference on Internet measurement*, pages 29–42. ACM, 2007.

[18] Mark EJ Newman. Finding community structure in networks using the eigenvectors of matrices. *Physical review E*, 74(3):036104, 2006.

[19] Gerard Roma, Perfecto Herrera, Massimiliano Zanin, Sergio L Toral, Frederic Font, and Xavier Serra. Small world networks and creativity in audio clip sharing. *International Journal of Social Network Mining*, 1(1):112–127, 2012.

[20] Markus Schedl. On the use of microblogging posts for similarity estimation and artist labeling. In *International Conference on Music Information Retrieval (ISMIR)*, pages 447–452, 2010.

[21] Markus Schedl. Analyzing the potential of microblogs for spatio-temporal popularity estimation of music artists. In *In Proceedings of the IJCAI 2011: International workshop on social web mining*, 2011.

[22] Karsten Steinhaeuser and Nitesh V Chawla. Identifying and evaluating community structure in complex networks. *Pattern Recognition Letters*, 31(5):413–421, 2010.