SONG RANKING BASED ON PIRACY IN PEER-TO-PEER NETWORKS

Noam Koenigstein **Yuval Shavitt**
School of Electrical Engineering
Tel Aviv University, Israel
{noamk, shavitt}@eng.tau.ac.il

ABSTRACT

Music sales are loosing their role as a means for music dissemination but are still used by the music industry for ranking artist success, e.g., in the Billboard Magazine chart. Thus, it was suggested recently to use social networks as an alternative ranking system; a suggestion which is problematic due to the ease of manipulating the list and the difficulty of implementation. In this work we suggest to use logs of queries from peer-to-peer file-sharing systems for ranking song success. We show that the trend and fluctuations of the popularity of a song in the Billboard list have strong correlation (0.89) to the ones in a list built from the P2P network, and that the P2P list has a week advantage over the Billboard list. Namely, music sales are strongly correlated with music piracy.

1. INTRODUCTION

Peer-to-peer (P2P) networks are one of the internet's most popular applications. The number of users and traffic, is growing dramatically from year to year. Despite several recent high profile legal cases against P2P vendors and users, it seems that the P2P community at large remains strong and healthy. In fact, P2P networks gain more acceptance as many companies and organizations distribute software and updates via networks such as BitTorrent to save bandwidth (e.g., Ubuntu).

Some studies suggest that music piracy might increase legal sales [1, 2], and copyright owners are advised to start developing business models that will allow them to generate revenue from P2P activity. Pioneering suggestions to utilize P2P networks for the benefit of the music industry were made by Bhattacharjee *et al.* [3,4], where P2P activity was used to predict an album's life cycle on the Billboard's top 200 albums chart.

In our previous work [5] we showed how P2P queries can be used for early detecting unknown emerging artists. In this study we take a different approach; we suggest an alternative songs ranking based on file sharing activity, that might replace traditional artists ranking such as the Billboard. We measured music piracy using a data set of geographically identified P2P query string, and compared it to songs ranking on the Billborad Hot 100, which measures sales and air plays. We compiled popularity charts based on P2P activity, and show a strong correlation between music piracy and legal sales and air plays. We argue that ranking songs through measurement of P2P queries is a good predictor of peoples' taste, and has many advantages over other means of popularity ranking, which were suggested in the past, most notably using social networks [6].

The remainder of the paper is organized as follows: In Section 2 we introduce the data set used in this study, and the methodology used to collect it. In Section 3 we focus on comparing song popularity in P2P networks with their ranking on the Billboard. We discuss the significance of our finding and our conclusions in Section 4.

2. DATA-SETS AND METHODOLOGY

We use two data sources for this study:

- **P2P Search Queries**: A data-set of queries collected from the Gnutella file-sharing network over twenty three weeks from January the 7th 2007 to June 8th 2007.
- **The Billboard Hot 100** The Billboad Hot 100 weekly charts for 2007 as published by the Billboard Magazine.

These two data-sets were collected independently, yet this study reveals a strong relationship between them. However, before analyzing the commonalities and differences, let us first describe the data sets and the methodology used to collect them.

2.1 P2P Search Queries

Queries in a file sharing network represent their users current taste and interests. A query is issued upon a request by a user searching for a specific file, or content relevant to the search string. In this study we used data collected from the Gnutella network using the Skyrider systems [1]. This data-set and the technical details of the methodology used to collect it are described in more depth in [7].

[1] Skyrider was a startup company that developed file sharing applications and services. It has recently been closed down. The data-set was collected when the company was still active, and is available for academic research.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page.
© 2009 International Society for Music Information Retrieval.

2.1.1 The Gnutella File Sharing Network

In a study performed by Slyck.com, a website which tracks the number of users of different P2P applications, Gnutella was found among the three most popular P2P file-sharing applications together with eDonkey and FastTrack [8]. Furthermore, according to [9], Gnutella is the most popular file sharing network in the Internet today with a market share of more than 40%. It is mainly used for piracy of music. In [5] the top 500 most popular queries were manually classified, and it was found that 68% of the queries were music related. Together with adult content (22%), these two categories dominate the query traffic, accounting together for 90% of the queries. Gnutella is also among the most studied P2P networks in the literature [5, 7, 10–16].

2.1.2 Methodology

A query's origin IP address is required for its geographical classification according to its country of origin. While it is possible to capture a large quantity of Gnutella queries by deploying several hundred ultrapeer nodes [2] , it will not be possible to tell the origin IP address of most of these captured queries. The basic problem in identifying the origin of captured queries is that queries do not in general carry information regarding their origin. What they do usually carry is an "Out Of Band" (OOB) return IP address. This address allows clients that have content matching a query to respond to a location close to the origin of the query, without having to backtrack the path taken by the query message. However, as most queries come from firewalled clients, in most cases the OOB address will belong to the ultrapeer connected to the query origin, acting as a proxy on behalf of the query originator. Deducting the missing origin IP address is not trivial. We resolved this problem by using a hop counting technique that is further explained in [7].

The vast majority of the Gnutella network is comprised of Limewire clients (80%-85%) and Bearshare clients (6%-10%) [10]. The Limewire client does not allow users to perform any kind of automatic or robotic queries. It does not allow queries with the SHA1 extension [3] , nor does it allow the automatic re-sending of queries. When it does send duplicate queries, it uses a constant Message ID which enables a simple removal of any duplication. By recording only queries originating from Limewire clients, we were able to significantly reduce the amount of duplications and automatic (non-human) queries, without losing too much of the traffic. Capturing only Limewire queries is an easy task as Limewire "signs" the message ID associated with each message it sends. This signature can be easily verified by the intercepting node, allowing it to ignore queries from all other clients.

[2] Ultrapeer nodes are special nodes that route search queries and responses for users connected to them

[3] SHA1 queries are queries in which only the hash key of a known file is sent without a string. This is useful when a client already started downloading and needs more sources.

Rank	String	Occurrences
1	adult	36,290
2	akon	23,468
3	lil wayne	12,518
4	beyonce	11,987
5	this is why i'm hot	10,746
6	justin timberlake	10,193
7	porn	9,144
8	don't matter	9,047
9	fergie	8,979
10	fall out boy	8,077

Table 1. P2P popularity chart for week 9 of 2007

2.2 Data Set Statistics

A daily log file of queries, typically contained 25-40 million record lines, each line consists of the query string, a date/time field, and the IP address of the node issuing the query. The origin country for each query was resolved using MaxMind commercial GeoIp database. Similarly to the Billboard charts, we wanted to concentrate on data originated from the United States. We thus removed all the non US queries reducing 55%-60% of the data records.

Our data-set comprised of query strings collected over a period of 23 weeks from January the 7th 2007 to June 8th 2007. The activity on the Gnutella networks increases by 20%-25% over the weekend [7]. We thus used weekly samples taken on a Saturday or a Sunday of every week of that period. The total number of US originated query strings processed in this study is **185,598,176**.

2.3 The Billboard Hot 100

The Billboard Hot 100 is the United States music industry standard singles popularity chart issued weekly by Billboard magazine [17]. Chart rankings are based on radio play and sales data collected 10 days before the chart is released. The ranking process does not take into account file sharing activity. A new chart is compiled and officially released to the public each Thursday. The chart is dated with the week number of the Saturday after, but in this study we used dates and week numbers according to the actual release date of the chart, and ignored the date issued by Billboard magazine. To simplify time tracking in this paper, we use week numbers instead of full date to chronologically order the Billboard charts and the weekly file sharing data we collected. For example, the Billboard chart which was released on Thursday January 11th 2007 (week number 2), was dated by billboard to January 20th (week 3) but by us to week number 2. The current top 50 singles are published weekly on the magazine website, while the full historical charts are available to on-line subscribers for a small fee. A statistical model of songs ranking in the Hot 100 chart can be found in [18].

3. CORRELATION OF TRENDS

As described above, the Billboard Hot 100 chart ranks songs relative to each other, and does not reveal the number of

sales or air-plays measured during that week. In order to compare it to our file-sharing data, we compiled our own weekly P2P popularity charts based on the popularity of search strings. We measured the popularity of each string by aggregating the number of appearances intercepted from a US based origin on that week. Table 1 depicts the top 10 positions of the P2P chart generated on week 9 of 2007 (sampled on March 1 2007).

Obviously, the P2P charts include many non music related strings. The string "adult" for example, was ranked number one on every chart we compiled. Unlike the Billboard charts, the P2P charts included also artists names (not only single titles), and sometimes even different variations of the same strings. In order to avoid inaccuracies, we looked only at the position of a song's exact title in the chart. To have high probability that the Billboard songs are ranked on our chart, we compiled truncated charts of the top 2000 strings each. A weekly log file contained on average 1.73 million different strings. Therefore, the top 2000 is approximately one thousandth of the entire P2P popularity chart. The top songs of our P2P chart were queried about 300,000 times per week in the USA. The songs at location 2000 were queried about 4,500 times. The number of queries per rank follows Zipf's law [7], thus changes in a rank position indicate strong shifts in popularity. When a song is no longer on the top 2000, it exits the P2P chart. This however, doesn't mean it is no longer being downloaded. Similarly when a single exits the Billboard Hot 100 chart, it doesn't mean it is not being played on the radio or sold in stores. Therefore, when considering the correlation of trends between the two charts, one should focus on the weeks where a song is ranked on both charts.

3.1 Correlation Measurements

We define $\overline{B_s}$ and $\overline{P_s}$ as the chart vectors representing the song s on the Billboard and P2P chart respectively.

$$\overline{B_s} = \{b_s(1), b_s(2), ..., b_s(23)\} \quad (1)$$

$$\overline{P_s} = \{p_s(1), p_s(2), ..., p_s(23)\} \quad (2)$$

Where $b_s(w)$ and $p_s(w)$ are the positions of song s on the Billboard and the P2P chart on week w respectively. If song s was not in the chart, we set its position to ∞ for that week. The *support* of a chart vector is the time range that the song was ranked in the chart. Namely where $b_s(w) < \infty$ or $p_s(w) < \infty$. The *joint support* of a song s is the time range in which it simultaneously ranked in both charts.

Fig. 1 depicts the chart vectors $\overline{B_s}$ and $\overline{P_s}$ for 6 different songs. The solid blue graph is the song's ranking on the Billboard Hot 100, while the dashed green graph is the song's ranking on the P2P chart. The horizontal axis (x-axis) depicts the date measured in week numbers in 2007. The song titles and performing artists are written above each graph. Note that lower parts of the graph represent higher position on the charts (i.e., the top of the chart is 1, while the last place is 100 or 2000). Looking at Fig. 1, one can easily notice the correlation between these two time series. This correlation is vivid not only in the general trend of the line, but also in minor trends and fluctuations.

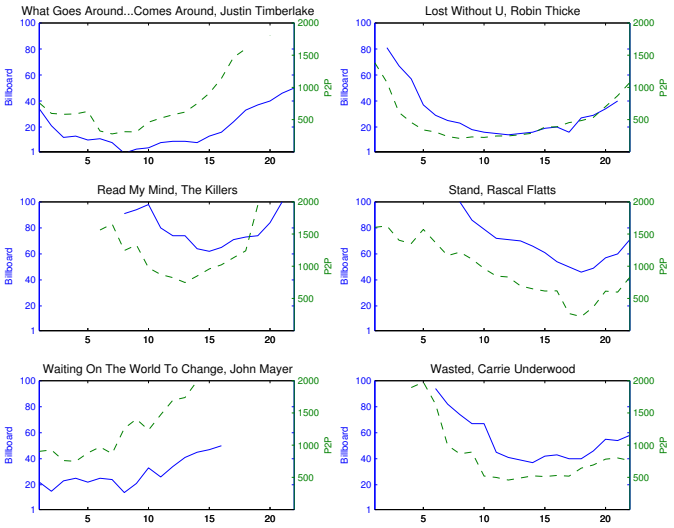


Figure 1. P2P Popularity Chart vs. The Billboard Hot 100

We slightly altered the standard definition of cross-correlation to consider only the joint support of the two series $\overline{B_s}$ and $\overline{P_s}$:

$$corr = \frac{\sum_{i=w_s}^{w_e} [(b_s(i) - E\{\overline{B_s}\}) \cdot (p_s(i) - E\{\overline{P_s}\})]}{\sqrt{\sum_{i=w_s}^{w_e} (b_s(i) - E\{\overline{B_s}\})^2} \sqrt{\sum_{i=w_s}^{w_e} (p_s(i) - E\{\overline{P_s}\})^2}} \quad (3)$$

Where $[w_s, w_{s+1}, ..., w_e]$ is the joint support and $E\{\overline{B_s}\}$ and $E\{\overline{P_s}\}$ are the means of the corresponding series. The correlation coefficient is in the range of $-1 \leq corr \leq 1$, where the bounds indicating exact match up to a scaling factor, while 0 indicates no correlation.

In all our measurements, we required songs to have a joint support of at least 4 weeks. This is the majority of the date-set (over 80%). Songs with a joint support of less than 4 weeks are mainly songs that ranked before or after our measurements, and had only a short "tail" inside our measurement period. Such songs poorly represent correlation of popularity trends over time.

We measured the correlation coefficients of the 135 songs that had a joint support of at least 4 weeks within the first twenty three weeks of 2007. The average joint support was 10.9 weeks. The average correlation coefficients was 0.67 while the median was 0.82, indicating a very strong correlation.

One might argue that the high correlation coefficients are the result of trend similarities of any time series of songs on charts. We thus measured the cross-correlation coefficient between the songs in one chart, and a random permutation in the other chart. Of the 52 songs which had a joint support of at least 4 weeks, the average joint support was 9.72 weeks, the average of the correlation coefficients was -0.006, and the median was 0.023, which negates the above hypothesis.

Title	Artist	No Shift	One Week
What Goes Around...Comes Around	Justin Timberlake	0.729	0.9707
Lost Without U	Robin Thicke	0.7664	0.948
Read My Mind	The Killers	0.1764	0.661
Stand	Rascal Flatts	0.9617	0.8965
Waiting On The World To Change	John Mayer	0.723	0.8965
Wasted	Carrie Underwood	0.8611	0.9456

Table 2. Correlation coefficients of the songs in Fig. 1

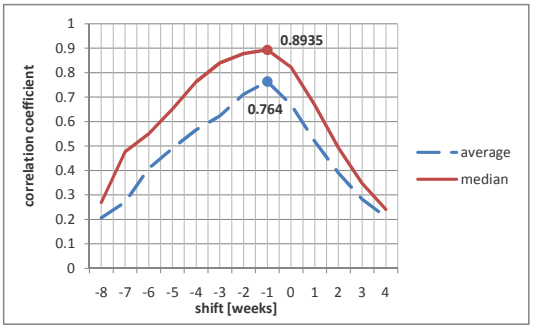


Figure 2. Cross-Correlation Coefficients vs. Time Shift

As mentioned is Section 2, the Billboard charts were dated according to their release date. However, the data used to compile each chart, is collected during the 10 days before the chart is published. Thus, we were interested in the correlation coefficient between the P2P chart and the Billboard chart of the following week. By shifting the Billboard chart vectors backwards, we measured the correlation coefficients of the 130 songs with a joint support of at least 4 weeks. The average joint support was 10.8 weeks. The average correlation coefficient was 0.76, while the median value was 0.89. These values are higher than the previous ones, which indicate a short time shift between the two series. Fig. 2 depicts the average and median values of the correlation coefficients, as a function of the Billboard's time shift. Clearly, minus one is the optimal time shift. We thus conclude that trends on the Billboard chart and on the the P2P charts are highly correlated with the Billboard lagging by one week. Table 2 depicts the correlation coefficients of the example songs in Fig. 1 without shifts, and with a one week time shift. When carefully examining Fig. 1, this time shift is noticed on some of the song graphs. The implication of this finding is obvious: P2P popularity charts can be used in order to predict trends on the Billboard chart. Record companies, for example, might use P2P file sharing activity to improve their marketing decisions.

3.2 Ranking Drift Analysis

In Section 3.1 we showed that songs trends (a climb or a descend) in P2P popularity charts are highly correlated with trends on the Billboard Hot 100. We now ask whether the charts are similar also in the relative ranks of songs. For each week we took the 100 songs from the Billboard chart, and "re-ranked" them according to their relative position on the P2P chart. In accordance with Section 3.1, we used a time shift of one week. We thus created an alternative ranking chart for the Billboard songs based on their P2P activity. This alternative chart is actually a filtered version of P2P chart from Section 3.1 that contains only the songs from the Billboard chart.

Rank	Billboard	Alternative Chart
1	*Irreplaceable*, Beyonce	*Walk It Out*, Unk
2	*I Wanna Love You*, Akon Feat. Snoop Dogg	*You*, Lloyd Feat. Lil Wayne
3	*Fergalicious*, Fergie	*Tim McGraw*, Tim McGraw With Faith Hill
4	*Smack That*, Akon Feat. Eminem	*Smack That*, Akon Feat. Eminem
5	*Say It Right*, Nelly Furtado	*We Fly High*, Jim Jones
6	*My Love*, Justin Timberlake Feat. T.I.	*Runaway Love*, Ludacris Feat. Mary J. Blige
7	*How To Save A Life*, The Fray	*Say It Right*, Nelly Furtado
8	*We Fly High*, Jim Jones	*Walk Away*, Paula DeAnda Feat. The DEY
9	*Welcome To The Black Parade*, My Chemical Romance	*Make It Rain*, Fat Joe Feat. Lil Wayne
10	*It Ends Tonight*, The All-American Rejects	*I Wanna Love You*, Akon Feat. Snoop Dogg

Table 3. Billboard's Top Ten Published on January 11th 2007 vs. The Alternative Chart

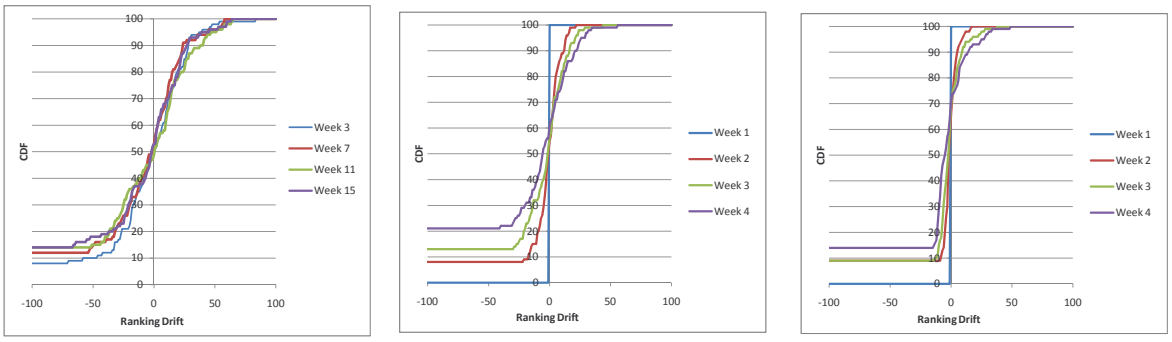
In Table 3 we show the top ten Billboard singles from the chart released on January 11th 2007 (week 2), and our alternative singles chart based on P2P activity on of the previous week. The two charts share four common songs, yet they are quite distinct. For the full 100 songs charts, the median distance of songs on the Billboard from their ranking on the alternative chart is 18.

In order to better understand the difference in song ranking, we define the *ranking drift* of a song as the difference between its rank on the Billboard chart to its rank on the corresponding alternative chart. We then plot the cumulative distribution function (CDF) of this difference for all 100 titles on the Billboard. Fig. 3(a) depicts the CDF of four weekly charts on different weeks in 2007. The week numbers are according to the Billboard charts. The correspondence of the two charts can be evaluated from the shape of graphs. A perfect match between the two charts, would appear as a perfect step function. Fig. 3(a) reveals a moderate correspondence of the Billboard charts with the alternative charts. For instance, the percentage of songs whose rank drift is in the range -25 to 25 is 60% on average.

CDF charts can be further used to compare the dynamics within each chart over time. We thus measured the drift of the songs from their ranking on previous weeks (on the same chart). Fig. 3(b) depicts the ranking drift of songs on the Billboard from the first week of 2007, over a period of 3 weeks. As expected the ranking drift increases for longer time intervals. Fig. 3(c) depicts the ranking drift of songs on the alternative chart during the same time period. Again the drift increases with time. The drift on the alternative chart, however, is smaller than that of the Billboard, indicating less change in songs ranking from week to week.

4. DISCUSSION

In past decades, air-plays and record sales were the primary means of distribution of popular music. The Billboard Hot 100 was therefore a reasonable proxy to popularity. Today, however, new technologies in particular the Internet, have created new means for distribution of music.



(a) Songs ranking drift between the Billboard Hot 100 and the alternative chart

(b) Songs time drift on the Billboard Hot 100

(c) Songs time drift on the alternative charts

Figure 3. Cumulative Distribution of Ranking Drift (CDF)

The growing popularity of file sharing make record sales and radio plays an increasingly poor predictor of peoples' taste. The record industry attempts to stop the swapping of pop music through the Internet by taking some P2P vendors to court, but the steady spread of file sharing systems and their technological improvements make them impossible to shut down.

In Section 3 we saw that currently the Billboard's sales based ranking system, is still quite in tune with what people download, but as file sharing becomes ever more prevalent, a need for a new ranking system arises. This observation was first introduced by Grace *et al.* [6], where it was suggested to use opinion mining (OM) on public boards to measure music popularity. In [6], comments on artists' pages on MySpace were used to build an alternative popularity chart of musical artists. Their top ten alternative list was substantially different than that of the Billboard. It was preferred, however, over the Billboard's list by 2-to-1 ratio by their 74 human test subjects.

We argue that popularity ranking based on P2P activity has many advantages over ranking based on opinion mining. First, it eliminates the complex task of classifying opinion polarities based on identifying opinion semantics. When P2P queries are considered, each query is always a positive indication of a user showing interest in the song or the artist. Second, the laborious task of identifying spam content in opinion mining, becomes trivial in a data set of query strings. On top of that, opinion mining in a website such as MySpace is biased towards the typical user of such a website, and biased again towards active users who care to comment on artists pages. Our method, doesn't require an active action on the side of the user. We rather measure queries generated as part of the file sharing process. Nonetheless, these queries disclose the interests of the user. Finally, we argue that opinion mining is more vulnerable to manipulations by stakeholders such as public relation companies acting on behalf of the artist or the record company. Planting comments on MySpace by interested entities is rather easy, while the technological barrier of generating many search queries in a file sharing network is much higher. In fact, networks such as Gnutella, already employ techniques to identify and eliminate non-human automatic search queries (as described in Section 1).

However, ranking songs based on P2P queries still has some open questions. There are, of course, the ethical issues with music piracy which are yet to be addressed. Regarding integrity, the ranking might be biased towards the preferences of file swappers which may differ in taste from the general public. It is also possible that a single P2P network, however large, has a user community which is biased against or for some genres, bringing the need to base the chart on all the top P2P networks and not just the largest one as was done in this study. Some of the open questions on the algorithmic side include the need to develop an artist ranking algorithm based on singles downloads, and to resolve the ranking of songs with confusing titles (e.g., *Love* or *Hot*).

It is not unlikely, that in the foreseeable future, music distribution based on file sharing will become the norm, and music sales will be reduced to a niche market. We expect that as the practice of file sharing becomes even more widespread, this line of research will become increasingly relevant.

5. REFERENCES

[1] Ram D. Gopal and G. Lawrence Sanders. Do artists benefit from online music sharing? *Journal of Business*, 79(3):1503–1534, May 2006.

[2] Martin Peitz and Patrick Waelbroeck. Why the music industry may gain from free downloading – the role of sampling. *International Journal of Industrial Organization*, 24(5):907–913, September 2006.

[3] Sudip Bhattacharjee, Ram D. Gopal, Kaveepan Lertwachara, and James R. Marsden. Using p2p sharing activity to improve business decision making: proof of concept for estimating product lifecycle. *Electronic Commerce Research and Applications*, 4(1):14–20, 2005.

[4] Sudip Bhattacharjee, Ram Gopal, Kaveepan Lertwachara, and James R. Marsden. Whatever happened to payola? an empirical analysis of online

music sharing. *Decis. Support Syst.*, 42(1):104–120, 2006.

[5] Noam Koenigstein, Yuval Shavitt, and Tomer Tankel. Spotting out emerging artists using geo-aware analysis of p2p query strings. In *The 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2008.

[6] J. Grace, D. Gruhl, K. Haas, M. Nagarajan, C. Robson, and N. Sahoo. Artist ranking through analysis of online community comments. *The 17th International World Wide Web Conference*, 2008.

[7] Adam Shaked Gish, Yuval Shavitt, and Tomer Tankel. Geographical statistics and characteristics of p2p query strings. In *The 6th International Workshop on Peer-to-Peer Systems*.

[8] Daniel Stutzbach, Reza Rejaie, and Subhabrata Sen. Characterizing unstructured overlay topologies in modern p2p file-sharing systems. *IEEE/ACM Transactions on Networking*, 16(2), 2008.

[9] Paul Resnikoff. Digital media desktop report, fourth quarter, 2007, April 2008. Digital Music Research Group.

[10] Amir H. Rasti, Daniel Stutzbach, and Reza Rejaie. On the long-term evolution of the two-tier gnutella overlay. In *IEEE Global Internet Symposium*, Barcelona, Spain, April 2006.

[11] Eytan Adar and Bernardo A. Huberman. Free riding on gnutella. *First Monday*, 5, 2000.

[12] M. Ripeanu. In *First International Conference on Peer-to-Peer Computing*.

[13] Matei Ripeanu, Ian Foster, and Adriana Iamnitchi. Mapping the gnutella network: Properties of large-scale peer-to-peer systems and implications for system design. *IEEE Internet Computing Journal*, 6:2002, 2002.

[14] A. Klemm, C. Lindemann, M. Vernon, and O. P. Waldhorst. Characterizing the query behavior in peer-to-peer file sharing systems. In *Internet Measurement Conference*.

[15] K. Sripanidkulchai. The popularity of gnutella queries and its implications on scalability, February 2001. Featured on O'Reilly's www.openp2p.com website.

[16] Mihajlo A. Jovanovic. Modeling large-scale peer-to-peer networks and a case study of gnutella. Master's thesis, University of Cincinatti, Cincinatti, OH, USA, 2001.

[17] Wikipedia the free encyclopedia. Billboard hot 100. http://en.wikipedia.org/wiki/Billboard_Hot_100 Last accessed May 2009.

[18] Eric T. Bradlow and Peter S. Fader. A bayesian lifetime model for the "hot 100" billboard songs. *Journal of the American Statistical Association*, 96:368–381, 2001.