Exploring J-DISC: Some Preliminary Analyses

Yun Hao
University of Illinois at Urbana-Champaign
yunhao2@illinois.edu

Kahyun Choi
University of Illinois at Urbana-Champaign
ckahyu2@illinois.edu

J. Stephen Downie
University of Illinois at Urbana-Champaign
jdownie@illinois.edu

ABSTRACT
J-DISC, a specialized digital library for information about jazz recording sessions that includes rich structured and searchable metadata, has the potential for supporting a wide range of studies on jazz, especially the musicological work of those interested in the social network aspects of jazz creation and production. This paper provides an overview of the entire J-DISC dataset. It also presents some exemplar analyses across this dataset to better illustrate the kinds of uses that musicologists could make of this collection. Our illustrative analyses include both informetric and network analyses of the entire J-DISC data which comprises data on 2,711 unique recording sessions associated with 3,744 distinct artists including such influential jazz figures as Dizzy Gillespie, Don Byas, Charlie Parker, John Coltrane and Kenny Dorham, etc. Our analyses also show that around 60% of the recording sessions included in J-DISC were recorded in New York City, Englewood Cliffs (NJ), Los Angeles (CA) and Paris during the year of 1923 to 2011. Furthermore, our analyses of the J-DISC data show the top venues captured in the J-DISC data include Rudy Van Gelder Studio, Birdland and Reeves Sound Studios. The potential research uses of the J-DISC data in both the DL (Digital Libraries) and MIR (Music Information Retrieval) domains are also briefly discussed.

Keywords
J-DISC; Jazz; Social Network; Metadata; Digital Libraries

1. INTRODUCTION
J-DISC is a specialized digital library for information about jazz recording sessions that includes rich structured and searchable metadata. Key entities in the data include *person, skill, session, track, composition,* and *issue*. Various relationships between the entities are recorded in 19 relational tables. Because of the extensive cultural, geographic, biographical, composer and studio information included, J-DISC has the potential for supporting a wide range of studies on jazz, especially the musicological work of those interested in the social network aspects of jazz creation and production. This paper presents findings from some illustrative analyses conducted on the entire J-DISC dataset, including informetric analyses in Section 2, and network analyses in Section 3. Concluding remarks are in Section 4.

The J-DISC dataset was created by the Center for Jazz Studies at Columbia University. J-DISC "…is organized to present complete information on jazz recording sessions, and merge a large corpus of session data into a single easily accessible repository, in a manner that can be easily searched, cross-searched, navigated and cited"[5]. The J-DISC team collected the data from a variety of vetted sources. The foundation of the collection is data on 75 "core" artists as determined by the J-DISC team in consultation with music scholars [4]. All the J-DISC records are stored in a Drupal data management system which is also used to create a searchable digital library (http://jdisc.columbia.edu). The J-DISC team created a unique metadata structure designed to better capture recording session-related data. Individual metadata records are available in XML for each session via the digital library. The J-DISC data schema is available at http://www.music-ir.org/mirex/gc15ux_jdisc/jdisc_schema.pdf. Information on how to access a copy of the data can be found at http://www.music-ir.org/mirex/wiki/2016:GC16UX:JDISC.

DLfM '16, August 12 2016, New York, USA

DOI: http://dx.doi.org/10.1145/2970044.2970050

2. INFORMETRIC ANALYSES
2.1 Sessions
A session is a social gathering of musicians and singers who perform music in a relatively informal context, and is an important part of jazz culture. In J-DISC, 2,711 unique recording sessions are represented using J-DISC's session-related metadata schema. The metadata include *name, venue (type, name, location), label, ensemble size, company, date, producer, engineer, sources,* and some other *details*. This section provides some illustrative descriptive informetrics derived from the J-DISC session metadata.

2.1.1 Ensemble Sizes
The number of artists involved in each J-DISC session ranges from 1 to 53. Figure 1 shows distribution of ensemble sizes in sessions. As is presented, about 41% of the sessions of which information about ensemble size is not null involve four or five artists; around 21% involve more than ten artists; more than 88% involve more than four artists. It can be concluded that sessions are not only musical events, but also social events from which interesting findings of social network of jazz artists might be reached.

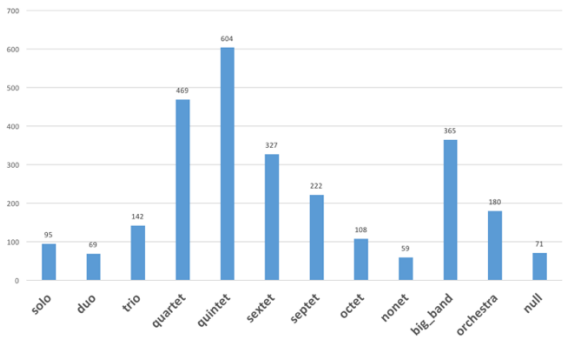

Figure 1. Ensemble sizes

2.1.2 *Time Span*

Figure 2 shows time span and distribution of the J-DISC recording sessions. The earlier sessions occurred around 1923 led by, for example, Jelly Roll Morton. The later sessions occurred around 2011 led by, for example, Paulo Moura and Gerald Wilson. Most of the sessions in J-DISC were recorded between 1940 to 1970.

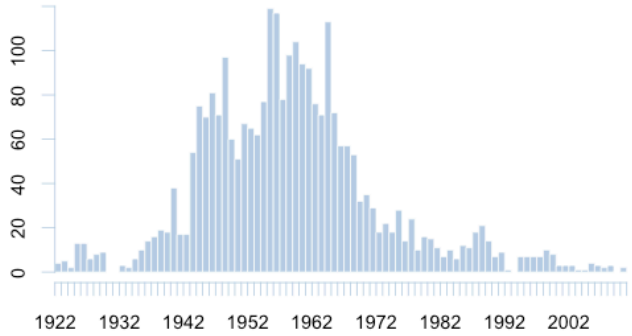

Figure 2. Time span of sessions

2.1.3 *Venues*

Venues where jazz music is played include clubs, dancehalls, theaters, studios, churches, etc. Most of the J-DISC venues are located in the United States and Europe. Several famous venues that appear the most often in the J-DISC data include Rudy Van Gelder Studio, Birdland, Reeves Sound Studios and Carnegie Hall. Figures 3 to 5 show distributions of venue locations in the J-DISC data. Figure 3 shows around 60% of the J-DISC recording sessions occurred in New York City, Englewood Cliffs (NJ), Los Angeles (CA) and Paris. Figure 4 shows that the J-DISC data predominantly represents jazz in the US. However, other countries such as France, Germany, and UK are also represented albeit to a lesser extent. Figure 5 shows a similar preponderance of New York-based sessions in the data. Figure 6 shows the change of venue types over time: studio sessions were the dominating type until mid-1960s. It is interesting to note the growth of live performance in the J-DISC data after the mid-1960s.

2.2 Artists

2.2.1 *Birth Dates and Gender*

There are 5,734 distinct artists and 3,774 artist associated with at least one session in the J-DISC data. J-DISC artists include such roles as composers, lyricists, producers, engineers of sessions, and performers and leaders in sessions. Most of the J-DISC artists were born between 1900 and 1955 (as shown in Figure 7). The gender distribution in the J-DISC artist data is overwhelmingly biased toward male artists: only 4% of the J-DISC artists are female, 93% male, and the remaining unknown (134 females, 3,487 males, and 123 unknown).

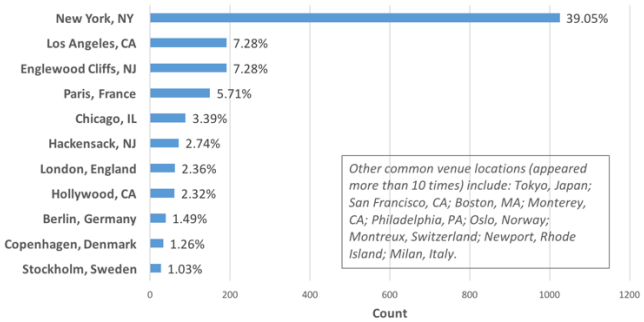

Figure 3. Venue locations

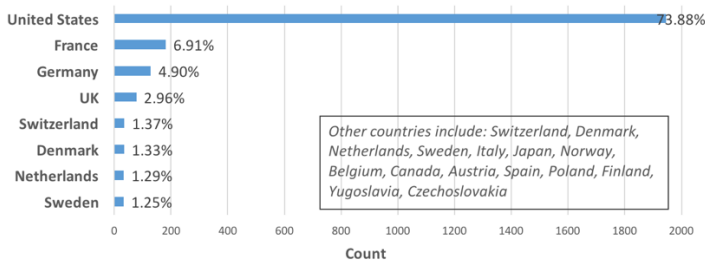

Figure 4. Venue locations by country

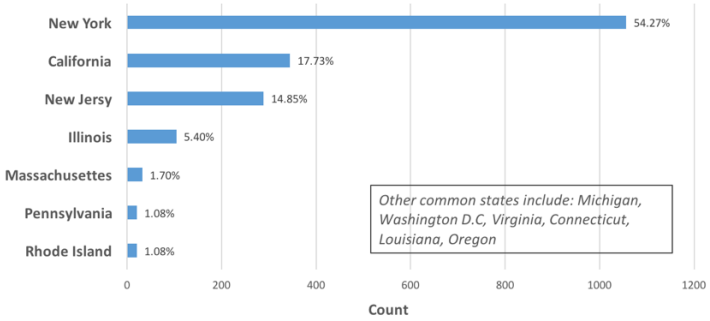

Figure 5. Venue locations in the US

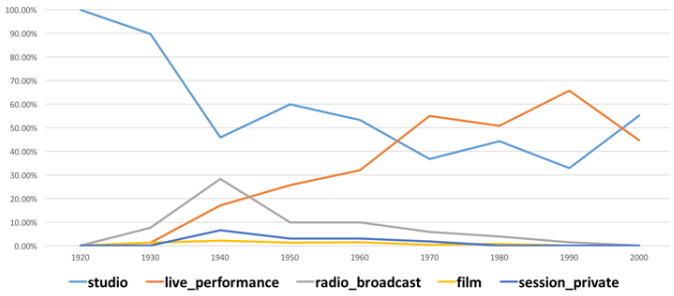

Figure 6. Change of venue type over time

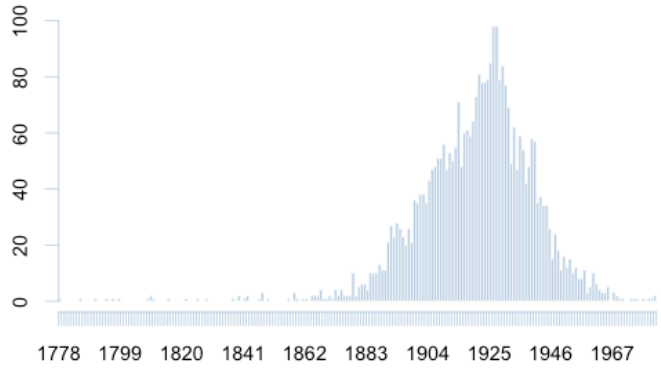

Figure 7. Birth dates of artists

2.2.2 *Top J-DISC Artists and Session Pairings*

The J-DISC data is useful for showing how jazz artists have interacted over the years. By summarizing the number of sessions in which each J-DISC artist participated and counting the number of sessions in which two artists performed together, a musicologist can illustrate the connectivity and possible influences among the artists. J-DISC has 21,424 artist entries that record an artist playing one or more instruments in a session, as well as 119,811 entries that record two artists playing together in a session. Table 1 shows the top 10 J-DISC artists who participated

in the most recording sessions. Table 2 shows the artists who performed in the most sessions together. Some of the network relationships among J-DISC artists performing together will be further discussed in Section 3.

Table 1. Top 10 most frequent artists involved in sessions

Rank	Artist	Cnt	Rank	Artist	Cnt
1	Dizzy Gillespie	432	**6**	Keith Jarrett	154
2	Don Byas	286	**7**	Billie Holiday	154
3	Cecil Taylor	260	**8**	Kenny Dorham	154
4	John Coltrane	245	**9**	James Moody	139
5	Charlie Parker	214	**10**	Wardell Gray	128

Table 2. Artists who frequently performed together

Rank	Artist_1	Artist_2	Cnt
1	John Coltrane	McCoy Tyner	85
2	Jimmy Lyons	Cecil Taylor	84
3	McCoy Tyner	Elvin Jones	82
4	John Coltrane	Jimmy Garrison	80
5	John Coltrane	Elvin Jones	77
6	Freddie Green	Count Basie	71
7	Dizzy Gillespie	James Moody	66
8	McCoy Tyner	Jimmy Garrison	63
9	Don Byas	Kenny Clarke	59
10	Jimmy Garrison	Elvin Jones	58

2.3 Instruments

There are around 150 instrument types represented in the J-DISC data. Figure 8 shows 10 of the most common instrument together with the number of sessions involving the instrument. As is shown, piano, drums and bass are the three most important instrument in the J-DISC sessions as they had been involved in around 90% of the sessions. Less than 20% of the J-DISC recording sessions involve vocal. In addition to the instruments shown in Figure 8, there are around 90 instrument types that appeared in fewer than 5 J-DISC sessions. On average, each session involves 6 different instrument types.

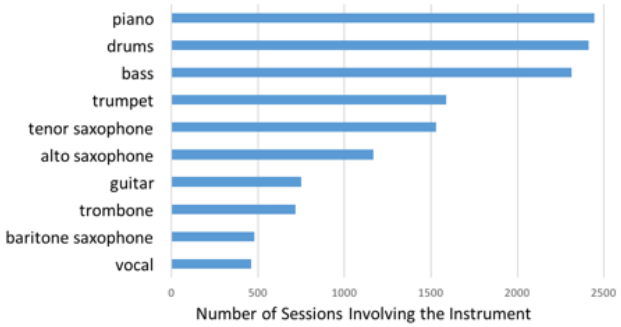

Figure 8. Common instruments

3. NETWORK ANALYSES

3.1 Social Networks

In the J-DISC data, there are 50 artists who performed with another artist in more than 30 different sessions. Table 2 shows the top 10 of these pairings. Space precludes showing the master network illustration of the top 50 artists. However, the master network consists of the six independent sub-networks presented in Figures 9 to 13. In the figures that follow, the relative thickness of an edge represents the relative frequency of the artists playing in the same session (i.e. the thicker, the more frequent). In Figure 9, 20 artists are connected with Dizzy Gillespie and Charlie Parker being the centers, showing evidence on the great influence of these two artists. This also reinforces the conclusion from [2] that in the 1940s Dizzy Gillespie, with Charlie Parker, became a major figure in the development of bebop and modern jazz.

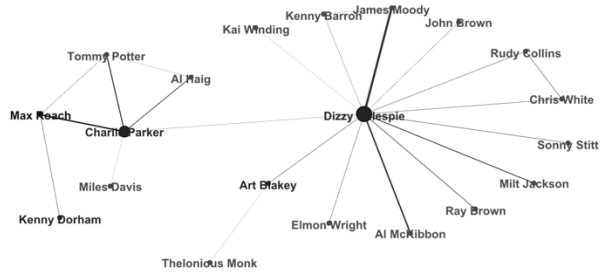

Figure 9. Sub-network featuring Gillespie and Parker

Figure 10 is another sub-network of 16 artists. Most of the artists in this figure had once joined Count Basie Orchestra, providing empirical evidence on the influence of Count Basie and his famous band.

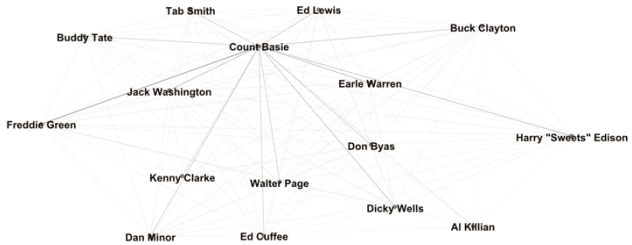

Figure 10. Sub-network featuring Basie

Figure 11 illustrates the 4 artists who most frequently performed with John Coltrane in the J-DISC dataset. In fact, Jimmy Garrison, McCoy Tyner and Elvin Jones were all members of John Coltrane's Classic Quartet. The edge between John Coltrane and Paul Chambers gives evidence of their frequent cooperation during the time when they both played in Miles Davis' band known as the "First Great Quintet".

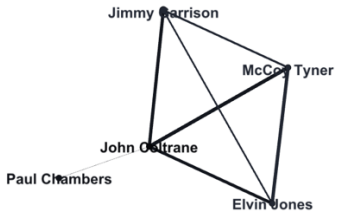

Figure 11. Sub-network featuring Coltrane

Figure 12 reinforces the long-lasting musical bond between Keith Jarrett and Jack DeJohnette during their careers, and provides evidence of cooperation between Benny Goodman and Wardell Gray in Goodman's bebop band bebop in the 1940s when bebop started to become popular.

Figure 12. Two sub-networks featuring Jarrett-DeJohnette, and Goodman-Gray

The social sub-network in Figure 13 featuring a heavily weighted edge between Jimmy Lyons and Cecil Taylor gives great evidence on the former's constant involvement in the Cecil Taylor Unit.

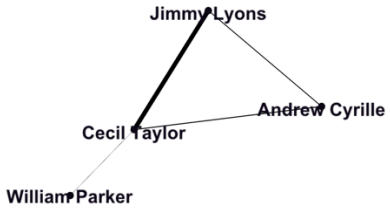

Figure 13. Sub-network featuring Lyons and Taylor

3.2 Betweenness, Eigenvector and Triangles

The J-DISC can also be used to calculate other kinds of network metrics. Such metrics include betweenness centrality, eigenvector centrality as well as the number of triangles with the artist being one vertex are calculated. Betweenness centrality and eigenvector centrality are calculated to discover the most salient artists (nodes): the nodes with high betweenness centrality act as hubs/bridges between other nodes. The eigenvector centrality can capture the influence of those nodes more than one hop away (global prominence) [3]. According to [1], if a graph is a social network, the number of triangles expected would be much greater than the value for a random graph. Therefore, the more frequently an artist is involved as a vertex of a triangle, the more connections that artist has with others. Tables 3 to 5 present respectively the top 10 names from the J-DISC data that are associated with the largest betweenness values, eigenvector values, and the greatest number of triangles. Note there is an overlap of four artists in the three lists; specifically, Dizzy Gillespie, Don Byas, Charlie Parker and Kenny Dorham appear in all three of the lists, with Dizzy Gillespie ranking first in all three. It is noteworthy that the remaining artists are unique to each list, demonstrating how each metric provides a different view of the data that might be of interest to musicologists.

Table 3. Top 10 artists with the largest betweenness centrality

R.	Artist	Btw.	R.	Artist	Btw.
1	Dizzy Gillespie	1481544	**6**	Billie Holiday	307284
2	Don Byas	995781	**7**	Wardell Gray	279149
3	Charlie Parker	480710	**8**	*Jelly Roll Morton*	236322
4	Cecil Taylor	386678	**9**	*Mary Lou Williams*	207149
5	Kenny Dorham	316248	**10**	*Lucky Thompson*	166597

Table 4. Top 10 artists with the largest eigenvector centrality

R.	Artist	Eig.	R	Artist	Eig.
1	Dizzy Gillespie	0.0085	**6**	*James Moody*	0.0032
2	Charlie Parker	0.0040	**7**	Wardell Gray	0.0031
3	Don Byas	0.0039	**8**	*Jerome Richardson*	0.0030
4	Kenny Dorham	0.0035	**9**	Billie Holiday	0.0030
5	*Clark Terry*	0.0032	**1**	Kenny Clarke	0.0027

Table 5. Top 10 artists with the most triangles

R.	Artist	# of Tri.	R.	Artist	# of Tri.
1	Dizzy Gillespie	911	**6**	Don Byas	288
2	*John Coltrane*	496	**7**	Charlie Parker	277
3	Cecil Taylor	385	**8**	Kenny Clarke	204
4	*Keith Jarrett*	332	**9**	*McCoy Tyner*	194
5	Kenny Dorham	310	**10**	*Elvin Jones*	181

4. CONCLUDING REMARKS

This paper provided an overview of the J-DISC dataset to demonstrate its unique value to potential research in both the DL and MIR domains. One of the powerful aspects of the J-DISC dataset is its richness in recording session-based data. Because "…the knowledge of the people, repertoire, and production and consumption processes involved in the domain of jazz recording can help in turn study or explore the professional and social networks, diverse subcultures, and artistic choices of jazz musicians across the history of the music"[6], J-DISC data can be of great value for musicological work especially those focused on social network aspects of jazz creation and production. Other potential uses of the J-DISC session-related metadata include building ground truth for MIR research. For example, building ground truth for artist similarity and instrument detection research, etc. Other uses include enabling new visualizations or developing interfaces for novel jazz music streaming services, etc. When associated with jazz audio, J-DISC data can also be used to create a comprehensive digital library and could be the foundation of new avenues of research of computational musicology. Although the data itself is limited in completeness, in the future J-DISC could still contribute to enhancing accessibility of jazz resources by becoming part of the Linked Open Data network to link to other jazz resources that are available online.

5. ACKNOWLEDGMENTS

We thank the Andrew W. Mellon Foundation and the J-DISC project for their financial and intellectual support.

6. REFERENCES

[1] Leskovec, J., Rajaraman, A. and Ullman, J.D., 2014. *Mining of massive datasets*. Cambridge University Press.

[2] Palmer, R., 2001. The Greatest Jazzman Of Them All? The Recorded Work of Dizzy Gillespie: An Appraisal. *Jazz Journal*, p.8.

[3] Wasserman, S. and Faust, K., 1994. *Social network analysis: Methods and applications* (Vol. 8). Cambridge university press.

[4] ITHAKA S+R, 2013. *J-DISC Sustainability Planning*, 13-14.

[5] Unique or distinctive features of jdisc.columbia.edu. Retrieved July 19, 2016 from Columbia University: http://jdisc.columbia.edu/content/unique-or-distinctive-features-jdisc/columbia.edu.

[6] The J-DISC Project: Background, Mission, and Recent Work. Retrieved July 19, 2016 from Jazz Studies Online: http://jazzstudiesonline.org/resource/j-disc-project-background-mission-and-recent-work.