Toward an Understanding of the History and Impact of User Studies in Music Information Retrieval

Jin Ha Lee

Information School, University of Washington, Mary Gates Hall, Ste 370, Seattle, WA 98195, USA

Phone: 206.685.0153; Fax: 206.616.3152; E-mail: jinhalee@uw.edu

Sally Jo Cunningham

Department of Computer Science, University of Waikato, Private Bag 3105, Hamilton 3240, New Zealand

Abstract Most Music Information Retrieval (MIR) researchers will agree that understanding users' needs and behaviors is critical for developing a good MIR system. The number of user studies in the MIR domain has been gradually increasing since the early 2000s, reflecting this growing appreciation of the need for empirical studies of users. However, despite the growing number of user studies and the wide recognition of their importance, it is unclear how great their impact has been in the field: on how systems are developed, how evaluation tasks are created, and how MIR system developers in particular understand critical concepts such as music similarity or music mood. In this paper, we present our analysis on the growth, publication and citation patterns, topics, and design of 198 user studies. This is followed by a discussion of a number of issues/challenges in conducting MIR user studies and distributing the research results. We conclude by making recommendations to increase the visibility and impact of user studies in the field.

Keywords *Music, MIR, User study, Citation analysis, Co-authorship analysis*

1 Introduction

Understanding users is a fundamental step in developing successful Music Information Retrieval (MIR) systems and services. Most MIR researchers will agree with this idea, and furthermore, it is not uncommon to hear speakers at MIR related conferences specifically arguing for the importance of user studies, academically as well as commercially. Despite the growing number of user studies and the wide recognition of their importance in the domain, it is unclear what impact these studies have really made. Have these studies in fact changed how MIR systems are developed or evaluation tasks are designed? Have they really changed how we understand critical concepts such as music similarity or mood? For MIR researchers specializing in user studies to move forward in this

domain, it is necessary to understand our past: what have we been doing and what kind of impact have we had? In order to lay the foundation for this discussion, we collected 198 user studies related to music, reviewed the content, and analyzed the research design and the publication and citation patterns of these studies.

2 Study Design

2.1 Definition of "User Studies"

Our first challenge was to define and set the boundaries for "user studies." From our analysis of relevant literature, we identified two major categories of user studies: "studies of users" (e.g., music information needs) and "studies involving users" (e.g., usability testing of MIR applications). Weigl and Guastavino [38], in their recent review article of user studies in MIR literature, defined user studies as "documents report(ing) on empirical investigations of user requirements or interactions with systems primarily aimed at providing access to musical information, including musical recordings, scores, lyrics, photography and artwork, and other associated metadata" (p. 335). In this study, we adopt a broader definition of "user studies" as studies reporting on 1) empirical investigation of needs, behaviors, perceptions, and opinions of humans through surveys, interviews, focus groups, and ethnographic methods, 2) experiments and usability testing involving humans focusing on a particular MIR system, 3) analysis of user-generated data, or 4) a summarization review of the studies above. Some may believe that studies under category 2 or 3 are not "true" user studies in the sense that the main focus in those studies can still be on the system rather than the users. However, we decided to adopt a broader definition as this will allow for a comparison of these different types of user studies and enable us to see patterns of concentration with regards to particular types of user studies related to MIR.

2.2 Data Collection

We conducted an extensive literature search in multiple domains related to music (e.g., MIR, Library and Information Science (LIS), Human-Computer Interaction (HCI), Computer Science (CS), Engineering, Psychology, Musicology) to identify these studies. We started our search by focusing on key venues for MIR studies such as the ISMIR proceedings, ACM DL, and IEEE Xplore. In conjunction with

the search keyword music, we used the following terms: user, human, people, need, use, behavior, information seeking, testing, involvement, learning, interaction, design, accessibility, usability, and user-centered. In order to be more precise in our search, we manually examined the title and abstract of the top retrieved results. After filtering the relevant studies, we also followed the citations in order to broaden our search. We ended up examining the publication metadata in multiple databases including WorldCat, Academic Search Complete (EBSCO), Web of Knowledge, and Google Scholar. In total, we found 198 studies related to music users between January 30 and February 2, 2013. We focused on the published academic/scientific literature and degree theses or dissertations. When our searches led to un-refereed or informally published work (e.g., white papers), we retained only those papers that had gained at least one citation in the research literature. We limited our search to include articles in journals, conferences, and workshops.

3 Publication Patterns of User Studies

3.1 Growth of the Publications

First, we analyzed several aspects related to the publication patterns of the user studies. We examined the publications dates of the user studies in order to learn more about the growth pattern. Figure 1 shows the distribution of the number of user studies published by year. We can observe the steady increase in the number of publications over the years. There were a small number of user studies pre-dating 2000, but the substantial growth started in early 2000s when the need for empirical user studies was pointed out in works such as [6], [10], and [14]. There was also a noticeable increase in 2009 although the growth pattern seems to have slowed down since 2010.

Figure 1. Distribution of the number of user studies by the year of publication

As a comparison, Figure 2 shows the change in the number of publications per year in the ISMIR proceedings. The figure, of course, only provides a limited view, not the change in the number of all publications related to MIR. However, considering that the ISMIR proceedings are a primary source of scholarly articles relating to MIR, this at least serves as a limited but important evidence for comparison. Although the growth pattern of user studies shown in Figure 1 is

encouraging, when compared with the number of studies focusing on the system aspect of MIR, the overall number of user studies still seems relatively small [38]. This trend is also evident in Figure 2, which shows the change in the number of publications along with the change in the number of publications related to user studies in ISMIR proceedings over the years. Although the number of publications in ISMIR proceedings has increased significantly over the past 13 years, the number of publications related to user studies in ISMIR proceedings has not increased at all (Mean = 3.5, Standard deviation = 0.27, Median = 3, Maximum 6, Minimum = 2).

Figure 2. Distribution of the number of publications in ISMIR proceedings per year

3.2 Publication Venues

We also examined the publication venues of these studies. Of the 198 studies, there were 126 conference publications, 63 journal articles, six workshop papers, one book chapter, and one white paper. There were a total of 92 different venues where music user studies appeared. The primary source of user studies was the ISMIR conference proceedings with 47 user studies followed by the CHI conference proceedings (12 user studies), and all the other journals and conference series each included six or fewer user studies. Sixty-eight of the 198 user studies (34%) were the only music user study published in that particular venue. Examples include several ACM and IEEE conference proceedings (e.g., ACM International Conference on Supporting Group Work; IEEE Computer Graphics & Applications), Psychology journals (e.g. British Journal of Psychology, Journal of Applied Psychology), LIS journals (Library & Information Science Research, Journal of Information Science), Music journals (e.g., Popular Music), etc. A few of them were conference proceedings and journals from countries outside of North America (e.g., Brazilian Symposium on Computer Graphics and Image Processing, Journal of the Korean Society for Information Management). This pattern of concentration in a small number of core publications can be explained by Bradford's law, which characterizes the pattern of diminishing returns in searching for references in scholarly publications [4]. Bradford's investigation of publication patterns in science journals revealed that the bulk of the papers on a special subject are published in a few journals specifically devoted to the subject [4]. He identified a pattern that "if scientific

journals are arranged in order of decreasing productivity of articles on a given subject, they may be divided into a nucleus of periodicals more particularly devoted to the subject and several other groups of zones containing the same number of articles as the nucleus" [11]. Bradford predicted a ratio of $1:n:n^2$ of journals where each proportion contains approximately the same number of articles. The concentration of MIR user studies in the ISMIR proceedings is perhaps stronger than Bradford's $1:n:n^2$ ratio of journals. Figure 3 shows the distribution of these studies by the titles of journals and conference proceedings. Only the titles that had 2 or more music user studies published in them were included in this figure.

Figure 3. Publication titles containing the user studies ordered by the number of studies

The wide distribution of publications across multiple domains/disciplines poses a challenge for researchers of user studies as well as readers who are interested in finding these studies. We confirmed that it is in fact impossible to find all these studies using a single database or search engine. Although it was possible to search for most of these articles on Google Scholar once we knew the title and the author information of the pertinent articles, conducting keyword searches using terms like "music" and "user" usually brought up a huge list of articles that contained many irrelevant ones, requiring a fairly long filtering process. Also since many publication venues only have small numbers of user studies, it is not very cost effective for users to conduct multiple searches in a number of databases just so that they can retrieve one or two more user studies. Another potential issue is that many researchers tend to conduct their literature search within their own domain, which will exclude many relevant works published in other fields (e.g., psychology scholars not citing MIR literature published in computer science venues). This pattern of individual researchers working within insular discourse communities, attending the same conferences and publishing primarily in the same journals has been well-discussed in information behavior literature; Dervin [8], for example, points out that the complexity and inaccessibility of literatures on human studies often results in researchers reinventing the wheel.

3.3 Co-authorship Analysis

We performed a co-authorship analysis to further understand the patterns of publication. Figure 4 shows the co-authorship graph generated by using NodeXL, a tool for visualization and exploration of networks [17]. The graph's vertices were grouped based on the Clauset-Newman-Moore cluster algorithm and the graph was laid out using the Harel-Koren Fast Multiscale layout algorithm. The nodes represent the authors and the line connecting the nodes represents the co-authorship between the two authors. The size of the node is scaled based on the number of publications by a particular author, and the width of the line connecting two nodes is scaled based on the number of times the pair of authors have co-authored a user study.

Figure 4. Co-authorship network among the authors of MIR user studies

A few strong networks emerged. The most notable network is grouped around Sally Jo Cunningham, J. Stephen Downie, Jin Ha Lee, David Bainbridge, Audrey Laplante, and 22 other scholars. Two additional networks are also very prominent: one formed around Jukka Holm and Arto Lehtiniemi, and another network formed around Charles Inskip, Andrew MacFarlane, and Pauline Rafferty. These strong networks seem to be based on specific labs/universities and regions: the University of Illinois, the University of Washington, and University of Waikato for the first group, Finland for the second group, and the UK for the third group. Another notable network formed around Adrian C. North and David J. Hargreaves in the UK represents many user studies published in psychology.

Another aspect to note is that the network is very disconnected, with a large number of small components, each consisting of a small number of authors. When we compare the network to the co-authorship network among all ISMIR authors generated in Lee et al. [22] (Fig. 5), it becomes evident that the fragmentation is much more severe in the co-authorship network among researchers of MIR user studies. In Figure 5, many US, Canada, and European research groups are tightly interconnected, resulting in a fairly large connected network in the middle. Part of the reason for this pattern could be that because MIR is still a relatively new field, there have not been many opportunities for cross-institutional ties to be formed. Fragmentation may reflect the fact that few researchers focus on this area, but larger numbers produce the occasional paper involving MIR user studies. Or, it may reflect the widespread appeal of music as a

subject for research (which is corroborated by the number and diversity of publication venues surveyed for this study). This pattern may also reflect the general co-authorship patterns of the larger IR community as mentioned in Smeaton et al. [36], who analyzed the past 25 years of publications from the SIGIR conference in 2002 and found that "the collaboration graph among SIGIR authors is very fragmented, with one large and several smaller components, as well as many disconnected authors or author pairs" although they did not provide statistics. The largest component of the SIGIR graph, however, seems to contain a much larger number of authors than our dataset (based on the analysis done by Hiemstra et al. [18]). They again analyzed the SIGIR proceedings from the past 30 years in 2007, and their co-author graph consisted of 361 different components with the main component containing 635 SIGIR authors (39%). Our graph, in contrast, consists of 88 different components with the biggest component containing 27 authors (8%). They also identified 1622 authors and 1150 papers as well as a growing trend in the average number of authors per paper. Conversely, our dataset shows a stable pattern of an average of 2 to 3 authors per paper since 2000.

As a comparison, Liu et al. [26] conducted a co-authorship analysis in the digital library (DL) research community (also a fairly interdisciplinary field). Their analysis was based on the publication data from the ACM DL, IEEE ADL (Advances in Digital Libraries), and JCDL (Joint Conference on Digital Libraries) conferences in which they identified 1567 authors, 759 publications, and 3410 co-authorship relationship pairs. Our dataset, on the other hand, contained 352 authors, 197 publications, and 413 co-authorship relationship pairs. The ratio between the number of authors and the co-authorship relationship pairs for the DL research is 0.46 compared to 0.85 in the MIR user studies, indicating the higher disconnect among the authors of MIR user studies.

Figure 5. Co-authorship network among ISMIR authors who have published two or more articles

We also generated separate co-authorship networks for different categories of user studies as defined in Section 2.1 (excluding the fourth category as the numbers of co-authorship pairs were minimal). Our definitions were as follows: 1) empirical investigation of needs, behaviors, perceptions, and opinions of humans through surveys, interviews, focus groups, and ethnographic methods, 2) experiments and usability testing involving humans focusing on a particular MIR

system, 3) analysis of user-generated data, or 4) review of the studies above. As Figure 6 shows, the first prominent networks are different for each of the category. For the first category of user studies, four strong networks can be identified: 1) one formed around Cunningham SJ and Bainbridge D, 2) MacFarlane A, Inskip C, and Rafferty P, 3) North AC and Hargreaves DJ, and 4) Downie JS, Lee JH, and Laplante A. However, for the second and the third categories of studies, for each only one dominant network stands out from a large number of smaller networks: for the second category, the network formed around Holm J and Letiniemi A is the most prominent, and for the third category, the network consisting of Lee JH, Downie JS, Cunningham SJ, and Bainbridge D is significantly more visible. We see from Figure 6 that only a small number of authors tend to repeatedly publish MIR user studies, and a large number of authors, especially the authors of studies based on user experiments and usability testing, tend to publish in this domain irregularly.

Figure 6. Co-authorship network among ISMIR authors of user studies in category 1, 2, and 3

Since the majority of the user studies were published after 2000, for a more complete picture, this analysis will have to be replicated in 10 or 20 years. Further analysis involving surveys and/or interviews with researchers of MIR user studies on how they find their co-authors and maintain the research ties may help determine the reasons for these kinds of co-authorship patterns.

4. Citation Patterns of User Studies

As part of the effort in understanding the impact of these studies, we investigated how often they were cited as of Feb 3, 2013 using the citation data from Google Scholar (GS). We used GS as a source of citation data because the major publications in the field (for example, the ISMIR conference proceedings) are not indexed in other major databases such as Web of Science, Scopus, etc. We found a total of 7504 citations of 198 user studies in research publications (two studies did not show up on GS and therefore we could not obtain the number of citing articles for those studies). Figure 7 shows the distribution of the citation counts for the user studies, where the X-axis represents the number of citations and the Y-axis represents the number of user studies that had the specified range of citation counts. The average number of citations was 38.3 with a standard deviation of 14.3, median of 8, and maximum of 2710. These statistics, however, are heavily

affected by the most cited article (Shardanand & Maes [34]), an outlier with 2710 citation counts. Excluding this article, the average number of citations was 24.6 with a standard deviation of 4.0, median of 8, and maximum of 425.

Figure 7. Distribution of the number of references of the user studies in other scholarly articles

Figure 8 shows the distribution of the citation counts of user studies by publication date. The X-axis represents the publication years of the cited user studies, and the Y-axis represents the number of publications citing the user studies. The figure shows a peak in 1995 (due to the most cited article by Shardanand & Maes [34]) and an inconsistent pattern, despite the growing trend in the number of user studies published as shown in Figure 1. The decrease in the number of references to user studies in later years may reflect the fact that some of them are yet to be cited because they are still recent. However, both Figure 7 and 8 also suggest that a small number of studies are driving up the overall citation counts.

Figure 8. Distribution of the number of references of user studies by their publication dates

Table 1 presents the 15 most cited user studies in the field. These include a mix of user experiments, evaluation of specific systems, studies of information behaviors and user-generated data, etc. The most heavily cited user study was by Shardanand & Maes [34], which describes a technique for making personalized user recommendations from any type of database based on similarities among user interest profiles. The article specifically discusses a system called Ringo, a personalized recommendation system for music albums and artists, and compares four different algorithms for making recommendations by using social information filtering. This was one of the earlier papers describing techniques for content-based, collaborative recommendation systems, and therefore was cited in many articles on recommendation systems for a wide range of cultural artifacts and commercial products, not limited to music.

The second most cited article was by Rentfrow & Gosling [32] on the structure and personality correlates of music preferences. The authors attempt to extend a theory of music preferences by examining individual differences and exploring the relationships between music preference and personality, self-views, and cognitive ability. Examining the citing articles on GS reveals that this particular work also has a broader impact outside of the music domain: it was not only heavily cited in research related to music preferences, but also emotion and

personality studies in social psychology as well. The high citation patterns of these two studies seem to be at least partially due to the interdisciplinary nature of the research, by inheriting higher citation counts from multiple fields. This effect of interdisciplinary nature of research on higher citation counts has been well documented in studies such as Levitt and Thelwall [25].

The third most cited article was by McNab et al. [27]. Ten users were asked to sing ten songs from memory, which were taped for analysis of key, pitch, contour, etc. The article was published in Proceedings of the First ACM International Conference on Digital Libraries and was cited widely in various papers on content-based music retrieval systems and measures. We believe that the heavy citation of this paper and also Levitin [24] was at least partly due to the fact that they were among the earliest papers focusing on content-based MIR, a topic which has dominated MIR research for the past decade.

Studies by Berenzweig et al. [2] and Ellis et al. [12] explore measures for generating ground truth based on user data, a topic strongly relevant to the evaluation of algorithms, another big accomplishment of the past decade (i.e., MIREX (Music Information Retrieval Evaluation eXchange)[1]). Studies of more general user needs and behaviors (North et al. [28], [29], Sloboda et al. [35], Lee & Downie [20], Cunningham et al. [7]) may have had a broader impact on multiple areas related to music. The popularity of the music application or device under study (e.g., Bull [5], Voida et al. [37]), the association between music and other multimedia (e.g., Boltz et al. [3]), and specific techniques for organizing music (Pauws & Eggen [30]) also seem associated with heavy citation patterns.

The publication venues of the top 15 most cited articles varied quite a bit and were from multiple domains. There were two articles from the CHI proceedings and the ISMIR Proceedings. Other titles were: Journal of Psychology and Social Psychology, Proceedings of the ACM International Conference on Digital Libraries, British Journal of Educational Psychology, Computer Music Journal, Perception & Psychophysics, Leisure Studies, Music Perception: An Interdisciplinary Journal, Journal of New Music Research, Memory & Cognition,

[1] Music Information Retrieval Evaluation eXhange (MIREX) is the annual evaluation campaign for various music information retrieval algorithms hosted by the International Music Information Retrieval Systems Evaluation Lab (IMIRSEL) at the University of Illinois at Urbana-Champaign.

Proceedings of the ACM/IEEE-CS Joint Conference on Digital Libraries, and Musicae Scientiae.

Table 1. The top 15 most cited user studies (excluding self-citations)

We also wanted to identify the most active researchers in the field, adopting the metric of the number of publications by each author. Table 2 shows the top 10 researchers who had published the largest number of user studies. It also provides the total number of citations of all publications by each author (excluding self-citations) as well as the number of co-authors for each author. Based on the venues where the user studies were published, we can infer that Cunningham, Lee, Downie, Inskip, Macfarlane, Bainbridge, Laplante, and Rafferty have been active in the core MIR field (e.g., published in the ISMIR Proceedings, ACM/IEEE Joint Conference on Digital Libraries, Journal of New Music Research), Holm and Lehtiniemi in mobile technology, multimedia, and visualization (e.g., published in the Proceedings of the International Conference on Mobile and Ubiquitous Multimedia, International Conference on Information Visualization), and North in music psychology (e.g., published in British Journal of Educational Psychology, Journal of Applied Social Psychology). The relatively high number of total citations for authors who are active in the MIR field is probably due to the fact that most user studies are published in this domain, followed by the music psychology domain. Sally Jo Cunningham had the highest number of co-authors as well as total citations and we can confirm this centrality in the co-authorship network as well (Fig. 3).

Table 2. The top 10 authors who published the highest number of user studies (excluding self-citations)

5 Topics of the User Studies

In addition to exploring publication and citation patterns, we conducted a topic analysis in order to gain an overview of the different topics that researchers have explored. For this analysis, we extracted all the title terms of the 198 publications in our user studies dataset. We stemmed the terms using a Perl-based implementation of the Porter stemming algorithm [31] and conventional stop words were removed. Table 3 shows the top terms (uni- and bigrams) that appeared in the publication titles of the user studies.

Table 3. Top 10 unigrams and bigrams from the title words of the user studies

We are able to infer several popular topics from these terms: digital music libraries, music recommendation systems, music preferences, music listening, evaluation, social network, and so on. Many of the top terms appear to represent the studies involving users to test and/or evaluate a system (e.g., retriev, system, evalu, retriev system). The verbs appearing in the list of unigrams suggest the studies have focused on music recommendations, interacting with music collections, and listening to music. Table 4 shows the top terms that appeared in the titles of the user studies for each year over the past 10 years. Common terms (i.e., music, user) were excluded. We highlighted the new terms entering the top-ranked lists in order to identify when certain topics first appeared in the user studies. The table shows that many title terms appeared quite recently (e.g., sharing, discovery, similarity in the past three years; recommendation, mood, tagging, context in the past five years) suggesting the continuing shift in the topics explored in the user studies.

Table 4. Top 15 ranked title terms over the past 10 years (w/ties); new terms are highlighted in bold.

In addition, we manually examined and extracted the topics covered in the analyzed articles, and used a card sorting technique in order to create a hierarchical taxonomy to represent the topics (Fig. 9). We identified three main categories that can be used to organize the sub-topics: Music, User, and System. Under each category, the topics were organized by their frequencies of appearances. In order to increase the impact of the user studies, we believe the explorations of a wider variety of topics need to continue and evolve. Although a wide variety of topics does not guarantee that the user studies will have high impact, it does mean they will have a broader impact, especially since MIR research trends do change over time [22].

Figure 9. Taxonomy of topics in MIR user studies

6 Research Design of User Studies

Lastly, we examined the studies more deeply in order to learn more about the research design of these user studies. The user studies employed a number of different research methods, which we analyzed to uncover the types and frequency of the various methods used (Fig. 10).

Figure 10. Research methods used in user studies

Experiments (46 studies), usability testing (42 studies), and interview (42 studies) were most commonly used methods, followed by survey (39 studies), and data mining/log analysis (34 studies). Other qualitative methods such as ethnographic observation (eight studies), diary study (eight studies), or focus group (four studies) were not used very often. Crowdsourcing (six studies) recently (from 2007) emerged as a new method and is becoming increasingly popular due to its speed and cost-effectiveness in data collection, evidenced by many studies discussed in [16]. Thirty-seven studies employed two or more methods, illustrating the increasing recognition of the utility of mixed method approaches.

The predominance of experiment and usability testing may suggest that we are heavily focusing on evaluating what is out there rather than focusing on deeper problems or questions, an issue noted in other areas such as HCI [15]. These studies are primarily evaluating performance (e.g., error rate/time to perform task with a new system); identifying usability issues (i.e., interface design problems); or investigating acceptability of new system/ interface. The full user-centered design process should include stages supporting coming to an understanding of the users, development of system prototype(s), and evaluation of the prototypes with users. However, relatively few papers presenting a new system include both an initial user requirements elicitation study and a follow-up performance/usability/acceptability study.

This heavy focus on experiment and usability testing seems to imply that many user studies assume the role of user to be passive rather than active in music listening and engagement. There seems to be an underlying assumption that users do not know what they want, how to describe their needs or requirements for particular music services/systems, or how to design such services/systems. Instead, those tasks are carried out by system developer/designers and users are asked to just provide feedback on the system that is already set up. Even in surveys or interviews, participants are simply asked to provide answers to given questions, with few opportunities for open-ended responses.

Another possible reason for this system-centric focus may lie in the grounding of the primary MIR publication venues (ISMIR, ACM and IEEE proceedings) in computer science and engineering: user-focused studies may not seem to reviewers to ‘fit’ with the conferences as neatly as system development and evaluation research. Studies based on ethnographic observation or diary study

might reveal more about the users' active music experience. However, as shown in the figure, they are used significantly less often than other research methods. This is not to say user studies employing qualitative methods such as ethnographic observation or diary study is superior in eliciting user information than usability testing or experiments. We believe the best approach, especially in studies involving humans, is to use multiple research methods and triangulate the research results in order to overcome the drawbacks of individual method.

We also investigated how many human subjects were involved in these studies. Tabulating the number of users involved in the study does not provide a full account on how much user effort is involved. However, it does help us get a general sense on the scale of these studies. One hundred and fifty-seven user studies involved human subjects, and 32 analyzed human-generated data such as queries, tags, etc. Nine studies did not directly involve human subjects or human-generated data, as they were papers based on literature review, meta-analysis, or theoretical reasoning. Figure 11 shows the distribution of the number of human subjects included in the studies of real users. Many studies are fairly small scale: 70 of the 157 studies (45%) involve 20 or fewer human subjects, and 123 studies (79%) involve 100 or fewer subjects. Most of the studies that have larger number of participants were from online surveys or surveys involving college students in Psychology departments.

Figure 11. Number of subjects in user studies

Note that the potential for active involvement of participants is limited with lab experiments and usability tests, which typically run at most a couple of hours. Ethnographic observations are constrained by the time available to the researcher to conduct observations. Interviews, surveys, and focus groups are attractive in that they may invite introspection and comment on music-related behavior over the long term, but at the cost of relying primarily on retrospection rather than direct, measurable experience. Only data mining/log analysis and the diary study naturally offer the opportunity to examine authentic music information behavior over the long term, though 'long term' studies are mainly of 1 to 4 weeks.

7 Discussion

Based on the results of our analyses as well as our own experiences in conducting music user studies, we provide a list of challenges/issues facing researchers who

conduct music user studies that require further discussion. We believe that these issues are stemming from the uniqueness of the subject and the research domain.

7.1 Fast-changing Field

We believe that the speed with which the MIR field has evolved has had a strong effect on both the scale of user studies as well as the longevity of the research findings of these studies. The rapid development of tools and technologies for music storage, distribution, and experience in the past few decades has been remarkable. Consider some of the most popular music related websites and applications today. Spotify was launched in late 2008 and in 2012, and has more than 10 million active users across Europe and the US. In their 2012 survey of 525 music users, Lee and Waterman [23] found that YouTube was the second most popular music related website. YouTube only launched in 2005, but it is now clearly affecting what users think of as the most useful music-related services. Music identification services and cloud music services, which are relatively new, are also increasingly becoming popular [23]. These developments illustrate that how our users envision and what they expect from music services are most likely changing rapidly as well. Most of the young adults today probably never had to deal with physical media and grew up with various music streaming services. The results from studies that investigated how people find and purchase music on such physical media will have limited applicability today.

We conjecture that the fast-changing field is at least one of the reasons for the prevalence of small-scale studies. Large-scale studies take longer, in terms of recruiting human subjects, as well as collecting and processing data, in particular if researchers want to incorporate a qualitative component. Longitudinal studies are by definition time-consuming. Due to the rapidly changing environment, researchers are constantly under pressure to conduct and publish studies swiftly. This can be especially true for those who are trying to test a particular system or methods for providing access to music, as there is a good chance that by the time the research gets published in a journal, the results are already outdated. This may also explain a large proportion of user studies being published in conference or workshop proceedings.

7.2 Issue of Generalization

A large proportion of MIR user studies are small to moderate scale studies investigating a limited number of users. With an exception of studies that employed data analysis of user search logs, queries, or reviews, only 22 of all the studies we identified were large-scale studies involving more than 100 human subjects. How does the scale of the study affect the generalizability of its results? Can we in fact make any reasonable inferences from studies of this scale that are generalizable to a larger user population? In addition, at least in certain parts of the world, it is not possible to obtain a comprehensive list of email addresses for surveying purposes, due to privacy concerns. This means that we often have to resort to convenience sampling, and study participants are in fact most frequently drawn from students or co-workers of the researchers—which again can negatively affect the generalizability of our findings.

A point worth noting here is that researchers of music users are trying to grapple with this nebulous idea of users. Who really are our users? In order to do a random sampling, by definition, we have to know the chance of a sample being selected from a population [1]. However, is it possible to clearly define the user population? Where do we draw the boundaries? Music is so pervasive in our lives that it is difficult to know who does and who does not interact with music. Moreover, music is often enjoyed and sought out across different regions and cultures. Many of the MIR systems/services are now being used by global user base. Thus researchers of music users, in some sense, are expected to derive findings that can potentially have global implications on a wide range of users across space and time. In this case, how do we define and randomly sample this population in a practical sense?

It is true that we can draw an artificial boundary and randomly sample a smaller population, say a large university in North America. Even when we do so, can we really say our sample is representative of our user group? In MIR studies, the subjects who respond to and participate in our studies will most likely be people who are interested in music to some degree. In this sense, the results are *always* likely to be biased. Some may not believe that this issue is problematic, considering that the people who are most likely to use the systems and services we develop are going to be those people who participate in our studies. However, this

discussion does illustrate the difficulty of defining the users and generalizing from results.

Due to these issues, we believe that rather than aiming for generalizing the research findings, it might be helpful to take an alternative approach to understanding the purpose of these studies: that each of these studies is discovering some piece of information about the users that is correct, but not comprehensive. When multiple pieces are put together, common themes emerge which we can generalize over multiple groups of users, as well as unique themes that can only apply to a particular user group.

7.3 Issues in Interpreting Study Results

MIR user studies are of course subject to common issues introducing an unintended bias to results, as identified in other fields. In evaluations of specific systems, the common finding is that the users like the new system and find the new interface entertaining or novel – but it is generally unclear how or whether participant behavior may change after the novelty effect wears off. Long term studies are desirable to move past the bias introduced by a novelty effect, but pressure to publish quickly in a fast-changing field (as well as the greater costs involved in conducting long term studies) encourages briefer studies. The predominance of complimentary evaluations of new systems may also reflect a publication bias towards positive results—a common bias found across many fields [13].

In addition, there is a good possibility that participants of these kinds of studies may be biased (consciously or unconsciously) to confirming the experimenter's hypotheses. In many of these studies, the participants are recruited from the same institution and sometimes the same department or lab. This may make participants feel that they have to focus more on the positive side of the system rather than sharing their true thoughts or opinion. This kind of user behavior has been well documented in psychology literature as the "Hawthorne effect" – the phenomenon in which subject in behavioral studies change their performance in response to being observed [9]. In order to alleviate this issue, it is crucial for researchers to make every effort to obtain more objective results by not asking leading questions, conducting blind testing, and concealing the intentions of the researchers. Further, given that most of the user studies researchers are

based in Europe and North America, the choice of participants does narrow the focus— which is evidenced by few studies of non-Western music users.

7.4 Lack of Systematic Synthesis of Research Results

Although a large proportion (26%) of user studies were published in the proceedings of ISMIR conference, other studies were published in journals and conferences in multiple domains including LIS, HCI, Music Digital Libraries (MDL), Multimedia Information Retrieval, Musicology, Psychology, etc. We had to repeat our search in multiple databases in order to retrieve the 198 studies scattered across multiple domains. Despite of our best efforts, we would not be surprised if there were additional studies we were not able to find—indeed, this potential lack of comprehensive cover is one of the main limitations of this study. Furthermore, although the ISMIR proceedings are freely available on the Web, a large number of other publications are fee-based. Unless the researchers' or readers' institutions have subscriptions to these different publications, it will be difficult and expensive to access these works. We suspect that this is probably one of the reasons hindering the synergic impact of these studies. Compared to more established domains such as Medicine or Law in which there are clearly dominant sources of scholarly articles (such as PubMed or LexisNesix, respectively), MIR is still a relatively new interdisciplinary field and there are multiple sources for scholarly articles from rather than one dominant source.

This also raises a question about distributing our knowledge to the general public--who are simply interested in music and not necessarily in the underlying research supporting their music access--and also to people who are in the music industry. Much of the MIR research aims to not only contribute to improving the general knowledge of music and how people interact with music, but also to create better systems and services related to music. If there is a barrier for general public and people outside of academia to access these works, then without a doubt, the impact we can make in the field will also be diminished. Although MIR researchers do disseminate their research results by using mediums other than scholarly articles such as radio, newspapers/magazines, blogs, and social media, access to the primary source will still be beneficial, as they will have the most complete description of the research results.

Without being able to easily find all the previous user studies that have dealt with similar research questions and user populations, we will essentially reinvent the wheel every time. This is in fact a common problem in all academic fields, but is further exacerbated in highly interdisciplinary research fields such as MIR. In order to resolve this issue, there is a need for additional review articles such as [38] and also an archive of all the citation information of user studies related to music. As the first step, we have made our initial set of paper references available as a searchable, browseable digital library[2], and are designing further enhancements to support community-driven contributions to the archive.

7.5 The Disconnect Between System/Evaluation Task Designers and User Studies Researchers

In the MIR domain, users' tastes or musical expertise are typically not taken into account in common evaluation strategies [33]. As an example, in the MIREX Audio Mood Classification task, the lyrical content is not considered in determining the ground truth although previous research (e.g., [21]) shows that lyrics do affect how people feel about music. For the Audio Music Similarity and Retrieval task, participants are simply asked to indicate how similar the given songs are based on their "musical similarity" without a clear definition of this concept [33]. Thus even though the task itself does involve human evaluators, what we have as a result of the evaluation may not be effectively translated to something meaningful or practical for real users. Developing this task into multiple sub-tasks such as a playlist generation task, known-item search task, or personal music collection management task seems closer to what would be useful for real users [19].

In MIREX, the evaluation task is typically proposed by researcher(s) who are involved in developing algorithms related to the task. In the MIR domain, however, researchers who conduct user studies are not always algorithm developers themselves; this is especially true for researchers engaged in studies of music users focusing on needs or behaviors. This disconnect may be one of the reasons why we have not seen a significant change in the way evaluation tasks have been run since MIREX started in 2005. Some of the suggestions made in the

[2] http://www.nzdl.org/greenstone3-mir-user-study/library/collection/mir-user-study/

user studies might be logistically difficult to implement, or the evaluators might not even agree with those suggestions. Without a more thorough investigation asking the system developers and the organizers of evaluation tasks, it will be premature to determine what the exact reasons are.

8 Conclusion and Future Work

In this paper, we reflected on how music user studies have been conducted and published, and what impact these studies have had on the field. Findings from our analysis suggest that there may be multiple layers of barriers for the user studies to make a strong impact: lack of findability due to the scattered patterns of publication, weak connections among scholars, dominance of small scaled studies that are difficult to generalize, and the disconnect between researchers conducting user studies and system developers/evaluation task designers.

This study is our first step in attempting to evaluate the impact of MIR user studies, and as such it of course is subject to limitations. First is the inevitable incompleteness of the list of user studies included in our analysis. In order to strive for precision, we chose to manually examine each user study to determine whether it should be included in our dataset rather than adopting an automatic approach based on the inclusion of particular keywords. Despite our best efforts, there is a limit on the number of articles we can examine as well as the number of databases we can search. We hope that by making our list of user studies publicly available, we are at least providing a starting point for developing a collaborative resource where researchers can continue to add new user studies. Another limitation is that some of the articles included in our dataset have been published by researchers who do not attend the MIR related conferences or publish in MIR related journals although they are well-cited in fields such as psychology, social science, etc. Therefore despite of the high citation counts, it is unclear how much impact these studies had on designing music services or MIR systems. We hope to address this issue in our future work in which we will aim to discover which studies were in fact deemed useful to MIR system developers and designers.

The main purpose of this work is to provide an opportunity for starting a discussion in the MIR community where many stakeholders involved in MIR research can together explore potential solutions to the issues raised in this paper.

Thus, we want to conclude our paper with a set of questions that need further discussion.

For researchers conducting user studies:

- How can we provide systematic and intelligent access to our work? Is there a sustainable method, perhaps a collaboratively managed resource?
- Is it necessary to change our research questions, methods, study populations, or venues in increase impact and affect change in the field?
- What kinds of difficulties/challenges exist in conducting the user studies?

For system and evaluation task designers/developers:

- How do you find out about new user studies and new study methodologies?
- Are there particular kinds of publications do you seek and use often?
- What kinds of user studies do you find most and least useful?
- What kinds of studies do you hope to see?

For everyone:

- How can we increase the collaboration between researchers conducting user studies and system developers/evaluation task designers?
- What do we see as the grand challenges in the area of MIR user studies?

In our future studies, we plan to survey and interview designers/developers of music related services and systems as well as organizers of MIREX evaluation tasks in order to more deeply understand the impact of these user studies. Specifically, we are interested in how the information on users are disseminated and diffused in the MIR domain, and how that knowledge may or may not affect the ways music services/systems are designed and modified. We will be asking questions related to how much awareness they have of the MIR research literature including the user studies. A deeper understanding on what kind of user information is actually sought by system designers/developers will be significant for researchers of MIR user studies.

Acknowledgments We thank Gary Gao, Tiffany Huang, and Ben Farabelli at University of Washington for their valuable contributions to this project, participants of the ISMIR 2012 Conference for their helpful feedback, David Bainbridge at the University of Waikato for creating

the initial Greenstone digital library of MIR user studies, and the anonymous reviewers for their comments and suggestions.

References

[1] Agresti A, Finlay B (2009) Statistical methods for the social sciences. Pearson Prentice Hall, Upper Saddle River, NJ

[2] Berenzweig A, Logan B, Ellis DPW, Whitman B (2004) A large-scale evaluation of acoustic and subjective music-similarity measures. Computer Music Journal 28(2): 63-76

[3] Boltz M, Schulkind M, Kantra S (1991) Effects of background music on the remembering of filmed events. Memory & Cognition 19(6): 593-606

[4] Bradford S (1985) Sources of information on specific subjects. Journal of Information Science 10(4): 173-180

[5] Bull M (2005) No dead air! The iPod and the culture of mobile listening. Leisure Studies 24(4): 343-355

[6] Cunningham SJ (2003) User studies: a first step in designing an MIR testbed. In: The MIR/MDL Evaluation Project White Paper Collection, GSLIS, Champaign, IL: 19-21

[7] Cunningham SJ, Reeves N, Britland M (2003) An ethnographic study of music information seeking: Implications for the design of a music digital library. Proceedings of the 3rd ACM/IEEE-CS Joint Conference on Digital Libraries: 5-16

[8] Dervin B (2003) Human studies and user studies: A call for methodological inter-disciplinarity. Information Research 9(1)

[9] Dickson WJ, Roethlisberger FJ (1966) Counseling in an organization: A sequel to the Hawthorn Researches. Harvard University, Boston, MA

[10] Downie JS (2003) Music information retrieval. In: Cronin B (ed) Annual Review of Information Science and Technology 37. Information Today, Medford, NJ, 295-340

[11] Drott MC (1981) Bradford's Law: Theory, Empiricism and the Gaps Between. Library Trends Summer (Special Issue on Bibliometrics): 41-52

[12] Ellis DPW, Whitman B, Berenzweig A, Lawrence S (2002) The quest for ground truth in musical artist similarity. Proceedings of the 3rd International Society for Music Information Retrieval (ISMIR) conference: 170-177

[13] Fanelli, D (2012) Negative results are disappearing from most disciplines and countries. Scientometrics, 90: 891-904

[14] Futrelle F, Downie JS (2002) Interdisciplinary communities and research issues in music information retrieval. Proceedings of the 3rd ISMIR Conference: 215-221

[15] Greenberg S, Buxton B (2008) Usability evaluation considered harmful (some of the time). Proceedings of the CHI'08: 111-120.

[16] Gomes C, Schneider D, Moraes K, De Souza J (2012) Crowdsourcing for music: survey and taxonomy. Proceedings of the IEEE SMCS Conference

[17] Hansen DL, Schneiderman B, Smith MA (2011) Analyzing social media networks with NodeXL: Insights from a connected world. Morgan Kaufmann, Burlington, MA

[18] Hiemstra D, Hauff C, De Jong F, Kraaij W (2007) SIGIR's 30th anniversary: An analysis of trends in IR research and the topology of its community. ACM SIGIR Forum, 41(2): 18-24

[19] Lee JH (2011) How similar is too similar?: Exploring users' perceptions of similarity in playlist evaluation. Proceedings of the 12th ISMIR Conference: 109-114

[20] Lee JH, Downie JS (2004) Survey of music information needs, uses, and seeking behaviours: Preliminary findings. Proceedings of the 5th ISMIR conference: 441-446

[21] Lee JH, Hill T, Work L (2012) What does music mood mean for real users? Proceedings of the 2012 iConference: 112-119

[22] Lee JH, Jones MC, Downie JS (2009) An analysis of ISMIR proceedings: Patterns of authorship, topic, and citation. Proceedings of the 10th ISMIR Conference: 57-62

[23] Lee JH, Waterman NM (2012) Understanding user requirements for music information services. Proceedings of the 13th ISMIR Conference: 253-258

[24] Levitin DJ (1994) Absolute memory for musical pitch: Evidence from the production of learned melodies. Perception & Psychophysics 54(4): 414-423

[25] Levitt JM, Thelwall M (2009) The most highly cited Library and Information Science articles: Interdisciplinary, first authors and citation patterns. Scientometrics 78(1): 45-67

[26] Liu X, Bollen J, Nelson ML, Van de Sompel, H (2005) Co-authorship networks in the digital library research community. Information Processing & Management 41(6): 1462-1480

[27] McNab RJ, Smith LA, Witten IH, Henderson CL, Cunningham SJ (1996) Towards the digital music library: tune retrieval from acoustic input. Proceedings of the 1st ACM International Conference on Digital Libraries: 11-18

[28] North AC, Hargreaves DJ, Hargreaves JJ (2004) Uses of music in everyday life. Music Perception: An Interdisciplinary Journal 22(1): 41-77

[29] North AC, Hargreaves DJ, O'Neill SA (2000) The importance of music to adolescents. British Journal of Educational Psychology 70: 255-272

[30] Pauws S, Eggen B (2002) PATS: Realization and user evaluation of an automatic playlist generator. Proceedings of the 3rd ISMIR Conference: 13-17

[31] Porter M (1997) An algorithm for suffix stripping. In: Sparck-Jones K, Willet P (eds) Readings in Information Retrieval. Morgan Kaufmann, San Francisco, CA: 313-316

[32] Rentfrow PJ and Gosling SD (2003) The Do Re Mi's of everyday life: The structure and personality correlates of music preference. Journal of Psychology and Social Psycology, 84(6): 1236-1256

[33] Schedl M, Flexer A (2012) Putting the user in the center of music information retrieval. Proceedings of the 13th ISMIR Conference: 385-390

[34] Shardanand U, Maes P (1995) Social information filtering: Algorithms for automating "word of mouth" Proceedings of the SIGCHI Conference on Human Factors in Computing Systems: 210-217

[35] Sloboda JA, O'Neill SA, Ivaldi A (2001) Functions of music in everyday life: An exploratory study using the Experience Sampling Method. Musicae Scientiae 5(1): 9-32

[36] Smeaton AF, Keogh G, Gurrin C, McDonald K, Sødring T (2002) Analysis of papers from twenty-five years of SIGIR conferences: What have we been doing for the last quarter of a century? ACM SIGIR Forum 37(1): 49-53

[37] Voida A, Grinter RE, Ducheneaut N, Edwards WK, Newman MW (2005) Listening in: Practices surrounding iTunes music sharing. CHI' 05 Proceedings of the SIGCHI Conference on Human Factors in Computing Systems: 191-200

[38] Weigl DM, Guastavino, C (2011) User studies in the music information retrieval literature. Proceedings of the 12th ISMIR Conference: 335-340

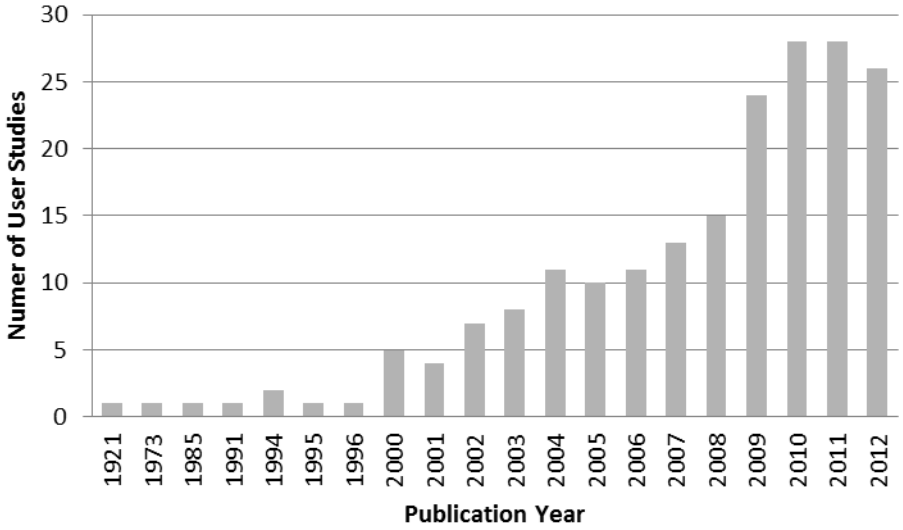


Figure 1. Distribution of the number of user studies by the year of publication

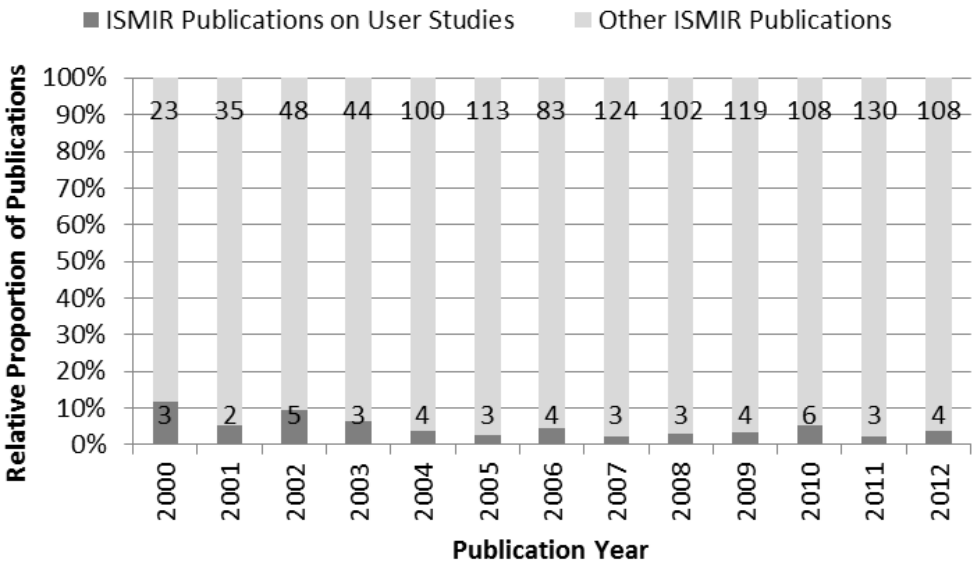
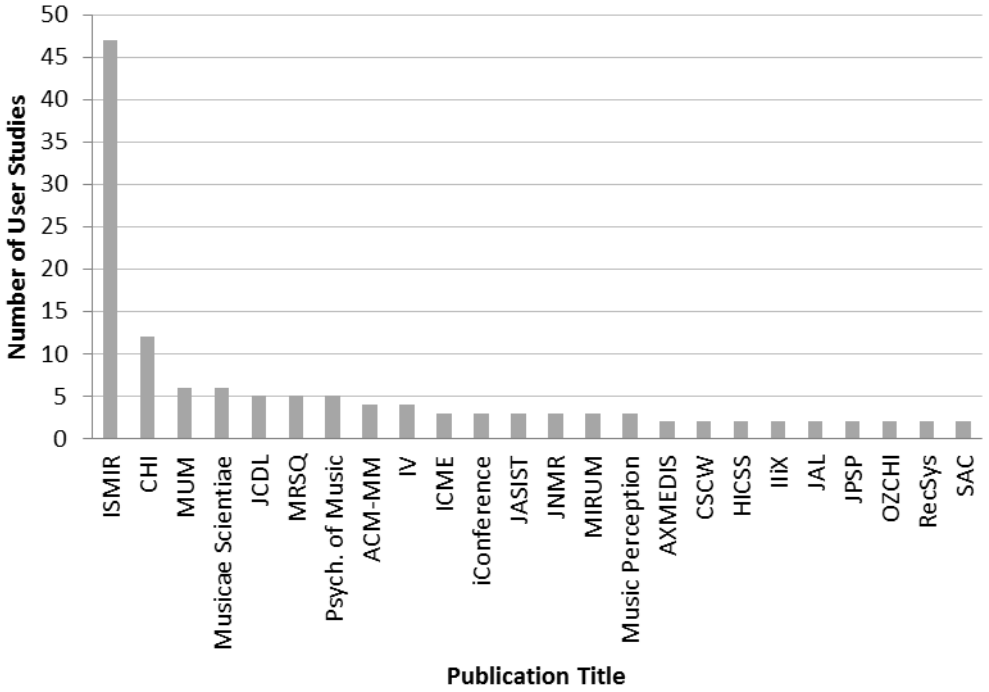


Figure 2. Relative proportion of user studies vs. non-user studies in ISMIR proceedings per year



* Full names of publications/conferences: ISMIR (International Society for Music Information Retrieval); CHI (ACM SIGCHI Conference on Human Factors in Computing Systems); MUM (International Conference on Mobile and Ubiquitous Multimedia); Musicae Scientiae; JCDL (ACM/IEEE-CS Joint Conference on Digital Libraries); MRSQ (Music Reference Services Quarterly); Psychology of Music; ACM-MM (International Conference on Multimedia); IV (International Conference on Information Visualization); ICME (IEEE International Conference on Multimedia and Expo); iConference; JASIST (Journal of the American Society for Information Science and Tehcnology); JNMR (Journal of New Music Research); MIRUM (International ACM Workshop on Music Information Retrieval with User-Centered and Multimodal Strategies); Music Perception: An Interdisciplinary Journal; Music Perception; AXMEDIS (International Conference on Automated Production of Cross Media Content for Multi-Channel Distribution); CSCW (ACM Conference on Computer Supported Cooperative Work and Social Computing); HICSS (Hawaii International Conference on System Sciences); IIiX (International Symposium on Information Interaction in Context);); JAL (Journal of Academic Librarianship); JPSP (Journal of Personality and Social Psychology); OZCHI (Australian Conference on Computer-Human Interaction); RecSys (ACM Conference on Recommender Systems); SAC (ACM Symposium on Applied Computing)

Figure 3. Publication titles containing the user studies ordered by the number of studies

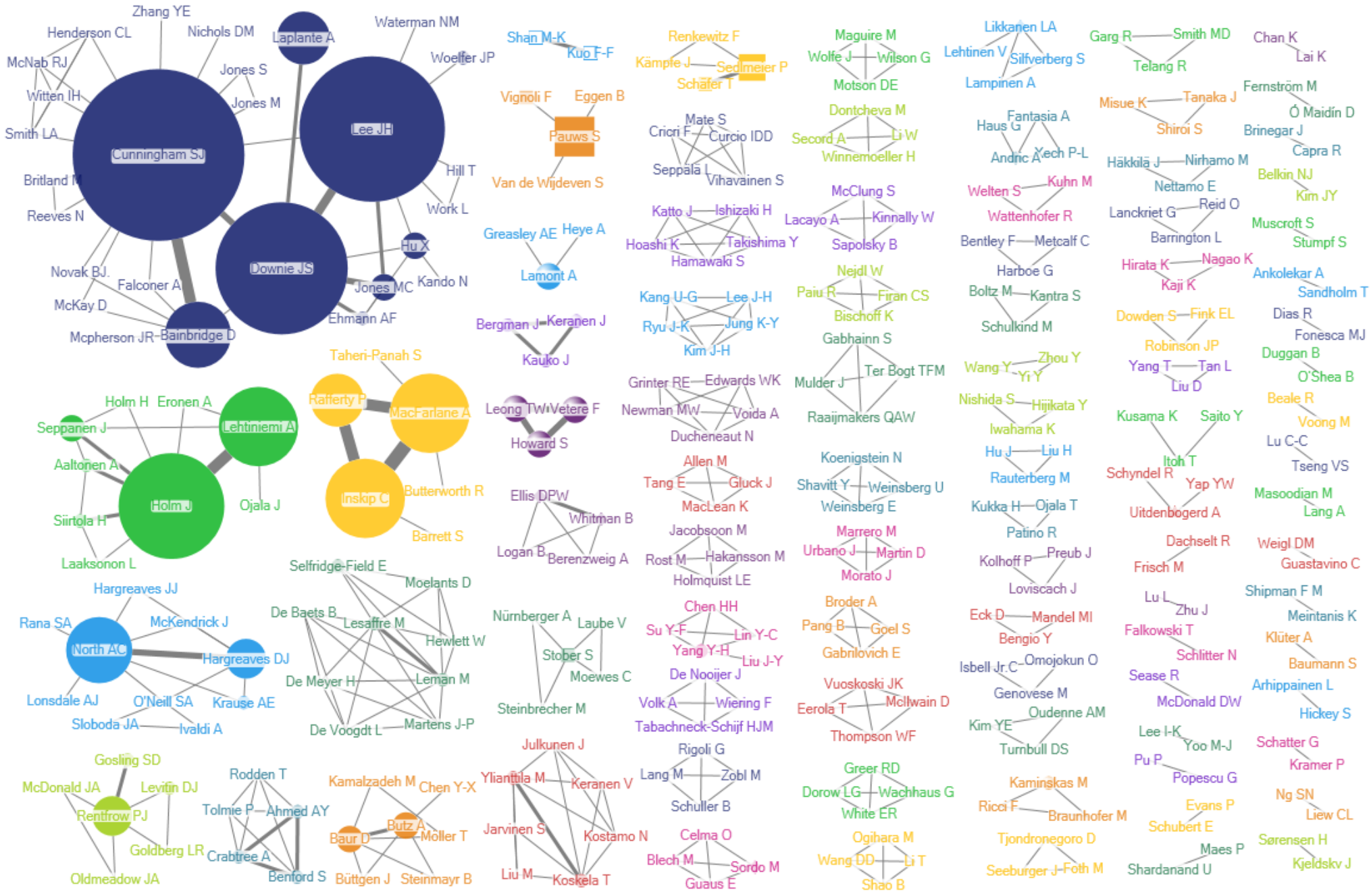

Zhang YE
Henderson CL
Nichols DM
Laplante A
Waterman NM
McNab RJ
Woelfer JP
Jones S
Witten IH
Jones M
Smith LA
Cunningham SJ
Lee JH
Hill T
Britland M
Work L
Reeves N
Hu X
Novak BJ
Downie JS
Falconer A
Jones MC
Kando N
McKay D
Ehmann AF
Mcpherson JR
Bainbridge D
Shan M-K
Kuo F-F
Vignoli F
Eggen B
Pauws S
Van de Wijdeven S
Greasley AE
Heye A
Lamont A
Bergman J
Keranen J
Kauko J
Renkewitz F
Kämpfe J
Sedlmeier P
Schäfer T
Mate S
Criori F
Curcio IDD
Seppälä L
Vihavainen S
Katto J
Ishizaki H
Hoashi K
Takishima Y
Hamawaki S
Kang U-G
Lee J-H
Ryu J-K
Jung K-Y
Kim J-H
Grinter RE
Edwards WK
Newman MW
Voida A
Ducheneaut N
Allen M
Tang E
Gluck J
MacLean K
Jacobsoon M
Rost M
Hakansson M
Holmquist LE
Chen HH
Su Y-F
Lin Y-C
Yang Y-H
Liu J-Y
De Nooijer J
Volk A
Wiering F
Tabachneck-Schijf HJM
Rigoli G
Lang M
Zobl M
Schuller B
Celma O
Blech M
Sordo M
Guaus E
Maguire M
Wolfe J
Wilson G
Motson DE
Dontcheva M
Secord A
Li W
Winnemoeller H
McClung S
Lacayo A
Kinnally W
Sapolsky B
Nejdl W
Paiu R
Firan CS
Bischoff K
Gabhainn S
Mulder J
Ter Bogt TFM
Raaijmakers QAW
Koenigstein N
Shavitt Y
Weinsberg U
Weinsberg E
Marrero M
Urbano J
Martin D
Morato J
Broder A
Pang B
Goel S
Gabrilovich E
Vuoskoski JK
Eerola T
McIlwain D
Thompson WF
Greer RD
Dorow LG
Wachhaus G
White ER
Ogihara M
Wang DD
Li T
Shao B
Likkanen LA
Lehtinen V
Silfverberg S
Lampinen A
Fantasia A
Haus G
Andric A
Xech P-L
Welten S
Kuhn M
Wattenhofer R
Bentley F
Metcalf C
Harboe G
Boltz M
Kantra S
Schulkind M
Wang Y
Zhou Y
Yi Y
Nishida S
Hijikata Y
Iwahama K
Hu J
Liu H
Rauterberg M
Kukka H
Ojala T
Patino R
Kolhoff P
Preub J
Loviscach J
Eck D
Mandel MI
Bengio Y
Isbell Jr.C
Omojokun O
Genovese M
Kim YE
Oudenne AM
Turnbull DS
Kaminskas M
Ricci F
Braunhofer M
Tjondronegoro D
Seeburger J
Foth M
Garg R
Smith MD
Telang R
Misue K
Tanaka J
Shiroi S
Häkkilä J
Nirhamo M
Nettamo E
Lanckriet G
Reid O
Barrington L
Hirata K
Nagao K
Kaji K
Dowden S
Fink EL
Robinson JP
Yang T
Tan L
Liu D
Kusama K
Saito Y
Itoh T
Schyndel R
Yap YW
Uitdenbogerd A
Dachselt R
Frisch M
Lu L
Zhu J
Falkowski T
Schlitter N
Sease R
McDonald DW
Lee I-K
Yoo M-J
Pu P
Popescu G
Evans P
Schubert E
Maes P
Shardanand U
Chan K
Lai K
Fernström M
Ó Maidín D
Brinegar J
Capra R
Belkin NJ
Kim JY
Muscroft S
Stumpf S
Ankolekar A
Sandholm T
Dias R
Fonesca MJ
Duggan B
O'Shea B
Beale R
Voong M
Lu C-C
Tseng VS
Masoodian M
Lang A
Weigl DM
Guastavino C
Shipman F M
Meintanis K
Klüter A
Baumann S
Arhippainen L
Hickey S
Schatter G
Kramer P
Ng SN
Liew CL
Sørensen H
Kjeldskv J
Holm H
Eronen A
Seppanen J
Lehtiniemi A
Aaltonen A
Holm J
Ojala J
Siirtola H
Laaksonon L
Taheri-Panah S
Rafferty P
MacFarlane A
Butterworth R
Inskip C
Barrett S
Leong TW
Vetere F
Howard S
Ellis DPW
Whitman B
Logan B
Berenzweig A
Selfridge-Field E
Moelants D
De Baets B
Lesaffre M
Hewlett W
De Meyer H
Leman M
De Voogdt L
Martens J-P
Nürnberger A
Laube V
Stober S
Moewes C
Steinbrecher M
Julkunen J
Ylianttila M
Keranen V
Jarvinen S
Kostamo N
Liu M
Koskela T
Hargreaves JJ
Rana SA
McKendrick J
North AC
Hargreaves DJ
Lonsdale AJ
O'Neill SA
Krause AE
Sloboda JA
Ivaldi A
Gosling SD
McDonald JA
Levitin DJ
Rentfrow PJ
Goldberg LR
Oldmeadow JA
Rodden T
Tolmie P
Ahmed AY
Crabtree A
Benford S
Kamalzadeh M
Chen Y-X
Butz A
Baur D
Moller T
Büttgen J
Steinmayr B

Figure 4. Co-authorship network among the authors

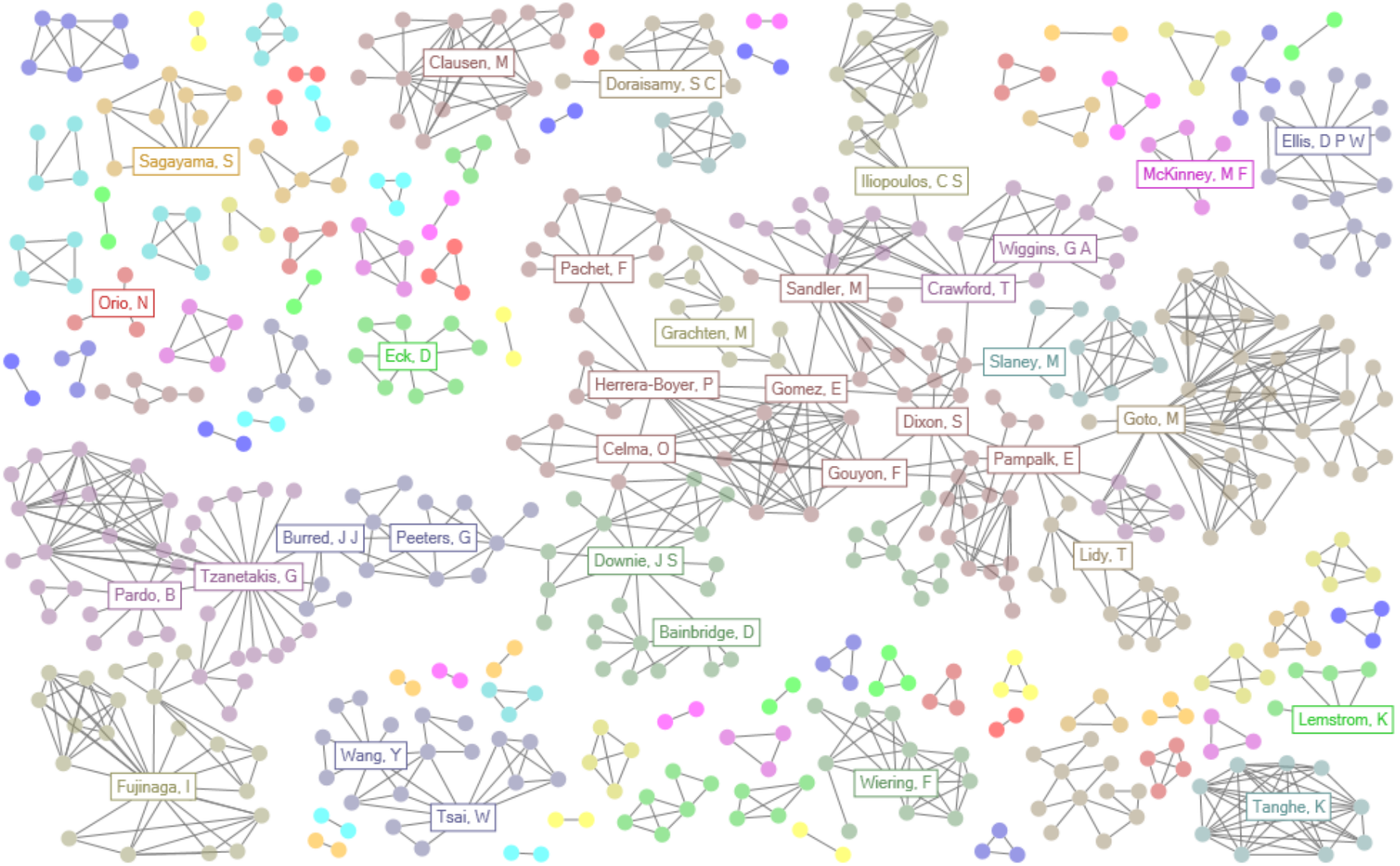
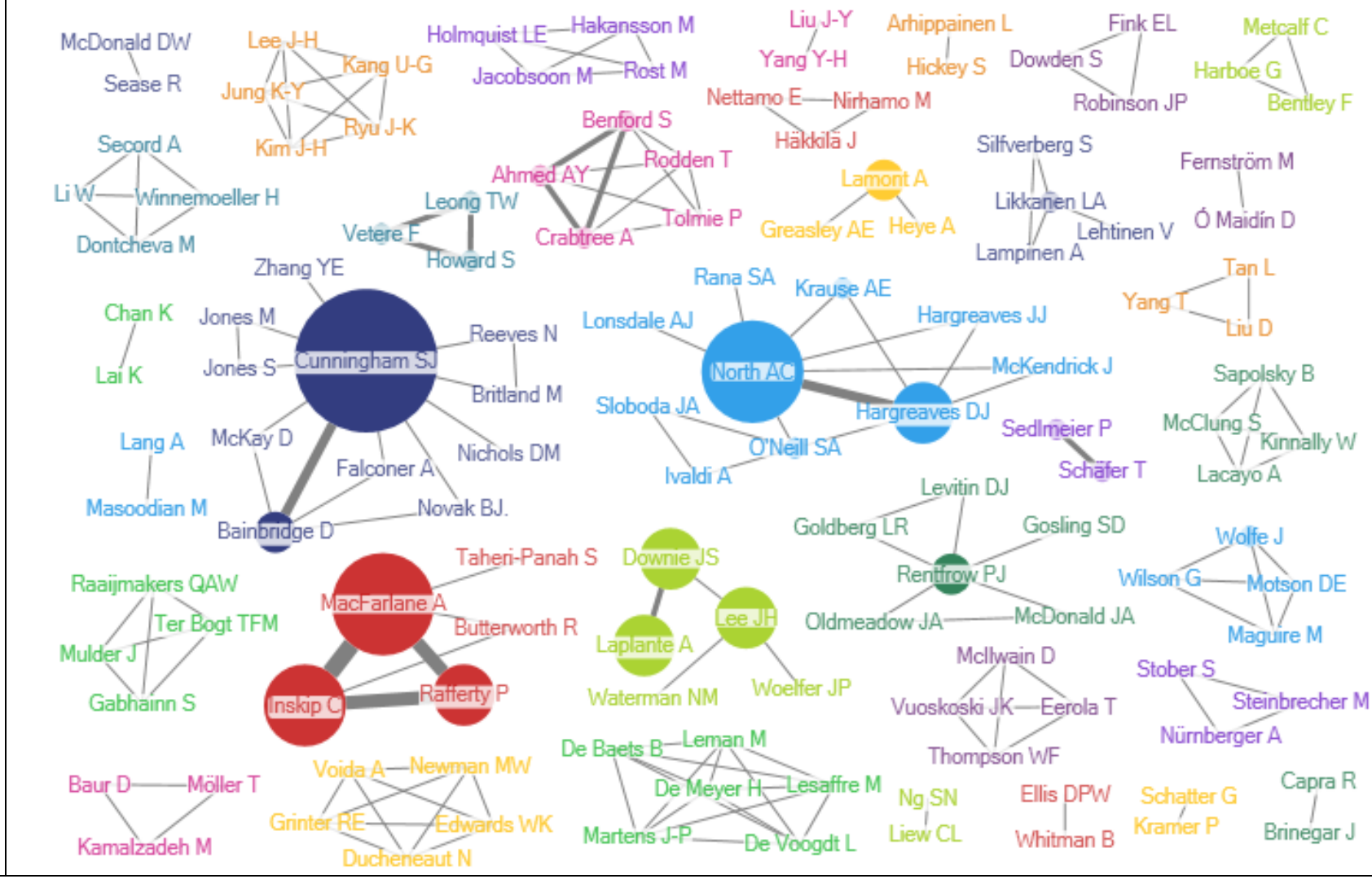
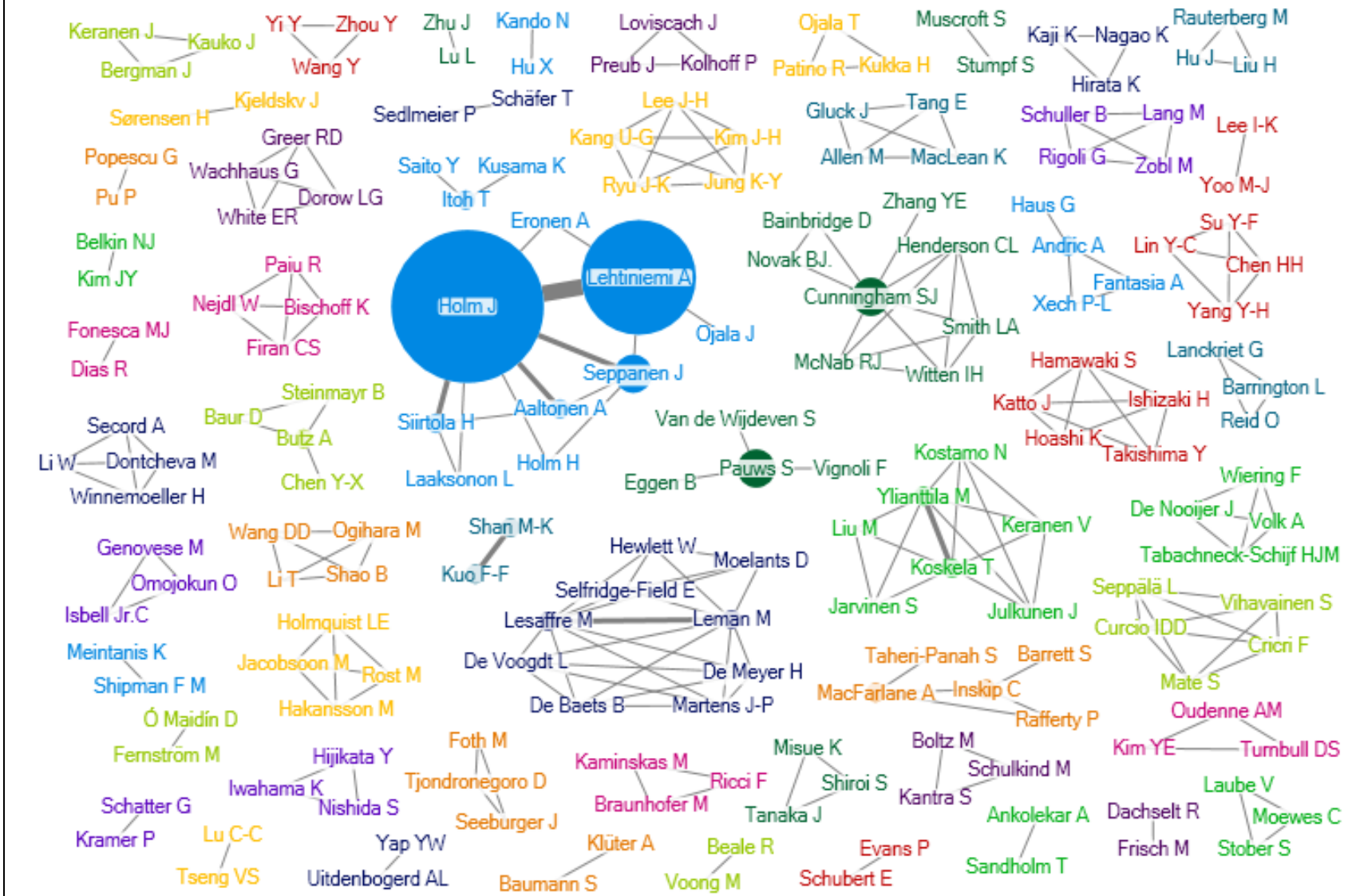

Clausen, M
Doraisamy, S C
Sagayama, S
Iliopoulos, C S
McKinney, M F
Ellis, D P W
Pachet, F
Wiggins, G A
Sandler, M
Crawford, T
Orio, N
Grachten, M
Eck, D
Slaney, M
Herrera-Boyer, P
Gomez, E
Dixon, S
Goto, M
Celma, O
Gouyon, F
Pampalk, E
Burred, J J
Peeters, G
Downie, J S
Lidy, T
Tzanetakis, G
Pardo, B
Bainbridge, D
Lemstrom, K
Wang, Y
Wiering, F
Fujinaga, I
Tsai, W
Tanghe, K

Figure 5. Co-authorship network among ISMIR authors who have published two or more articles (-2009)

Category 1


McDonald DW
Sease R
Lee J-H
Kang U-G
Jung K-Y
Ryu J-K
Kim J-H
Holmquist LE
Hakansson M
Jacobsoon M
Rost M
Liu J-Y
Yang Y-H
Arhippainen L
Hickey S
Fink EL
Dowden S
Robinson JP
Metcalf C
Harboe G
Bentley F
Nettamo E
Nirhamo M
Häkkilä J
Benford S
Rodden T
Ahmed AY
Tolmie P
Crabtree A
Leong TW
Vetere F
Howard S
Lamont A
Greasley AE
Heye A
Silfverberg S
Likkanen LA
Lehtinen V
Lampinen A
Fernström M
Ó Maidín D
Secord A
Li W
Winnemoeller H
Dontcheva M
Zhang YE
Chan K
Jones M
Reeves N
Jones S
Cunningham SJ
Britland M
Lai K
Lang A
McKay D
Nichols DM
Falconer A
Masoodian M
Novak BJ.
Bainbridge D
Rana SA
Krause AE
Lonsdale AJ
Hargreaves JJ
North AC
McKendrick J
Sloboda JA
Hargreaves DJ
O'Neill SA
Ivaldi A
Sedlmeier P
Schäfer T
Tan L
Yang T
Liu D
Sapolsky B
McClung S
Kinnally W
Lacayo A
Levitin DJ
Goldberg LR
Gosling SD
Renfrow PJ
Oldmeadow JA
McDonald JA
Wolfe J
Wilson G
Motson DE
Maguire M
Taheri-Panah S
Raaijmakers QAW
MacFarlane A
Butterworth R
Ter Bogt TFM
Mulder J
Gabhainn S
Inskip C
Rafferty P
Downie JS
Lee JH
Laplante A
Waterman NM
Woelfer JP
McIlwain D
Vuoskoski JK
Eerola T
Thompson WF
Stober S
Steinbrecher M
Nürnberger A
Baur D
Möller T
Kamalzadeh M
Voida A
Newman MW
Grinter RE
Edwards WK
Ducheneaut N
De Baets B
Leman M
De Meyer H
Lesaffre M
Martens J-P
De Voogdt L
Ng SN
Liew CL
Ellis DPW
Whitman B
Schatter G
Kramer P
Capra R
Brinegar J

Category 2


Keranen J
Kauko J
Bergman J
Yi Y
Zhou Y
Wang Y
Zhu J
Lu L
Kando N
Hu X
Loviscach J
Preub J
Kolhoff P
Ojala T
Patino R
Kukka H
Muscroft S
Stumpf S
Kaji K
Nagao K
Hirata K
Rauterberg M
Hu J
Liu H
Sørensen H
Kjeldsky J
Sedlmeier P
Schäfer T
Lee J-H
Kang U-G
Kim J-H
Ryu J-K
Jung K-Y
Gluck J
Tang E
Allen M
MacLean K
Schuller B
Lang M
Rigoli G
Zobl M
Lee I-K
Yoo M-J
Popescu G
Pu P
Greer RD
Wachhaus G
White ER
Dorow LG
Saito Y
Kusama K
Itoh T
Eronen A
Holm J
Lehtiniemi A
Ojala J
Sepppanen J
Aaltonen A
Siirtola H
Holm H
Laaksonon L
Zhang YE
Bainbridge D
Novak BJ.
Henderson CL
Cunningham SJ
Smith LA
McNab RJ
Witten IH
Haus G
Andric A
Xech P-L
Fantasia A
Su Y-F
Lin Y-C
Chen HH
Yang Y-H
Belkin NJ
Kim JY
Paiu R
Nejdl W
Bischoff K
Firan CS
Fonesca MJ
Dias R
Steinmayr B
Baur D
Butz A
Chen Y-X
Secord A
Li W
Dontcheva M
Winnemoeller H
Van de Wijdeven S
Eggen B
Pauws S
Vignoli F
Hamawaki S
Katto J
Ishizaki H
Hoashi K
Takishima Y
Lanckriet G
Barrington L
Reid O
Kostamo N
Ylianttila M
Liu M
Keranen V
Koskela T
Jarvinen S
Julkunen J
Wiering F
De Nooijer J
Volk A
Tabachneck-Schijf HJM
Wang DD
Ogihara M
Li T
Shao B
Shan M-K
Kuo F-F
Hewlett W
Moelants D
Selfridge-Field E
Lesaffre M
Leman M
De Voogdt L
De Meyer H
De Baets B
Martens J-P
Seppälä L
Vihavainen S
Curcio IDD
Cricri F
Mate S
Genovese M
Omojokun O
Isbell Jr.C
Holmquist LE
Jacobsoon M
Rost M
Hakansson M
Meintanis K
Shipman F M
Taheri-Panah S
Barrett S
MacFarlane A
Inskip C
Rafferty P
Ó Maidín D
Fernström M
Foth M
Tjondronegoro D
Seeburger J
Kaminskas M
Ricci F
Braunhofer M
Misue K
Shiroi S
Tanaka J
Boltz M
Schulkind M
Kantra S
Oudenne AM
Kim YE
Turnbull DS
Hijikata Y
Iwahama K
Nishida S
Schatter G
Kramer P
Lu C-C
Tseng VS
Yap YW
Uitdenbogerd AL
Klüter A
Baumann S
Beale R
Voong M
Evans P
Schubert E
Ankolekar A
Sandholm T
Dachselt R
Frisch M
Laube V
Moewes C
Stober S

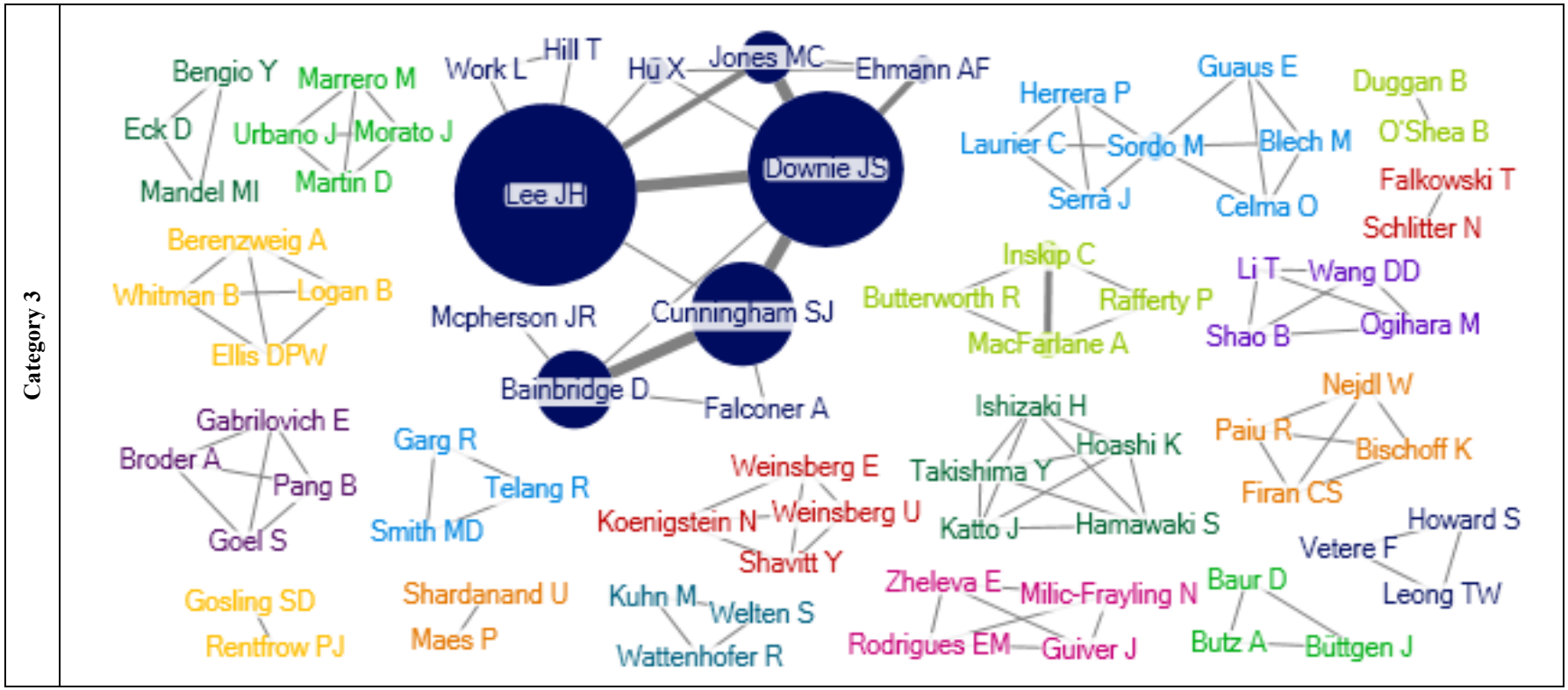


Figure 6. Co-authorship network among ISMIR authors of user studies in category 1, 2, and 3

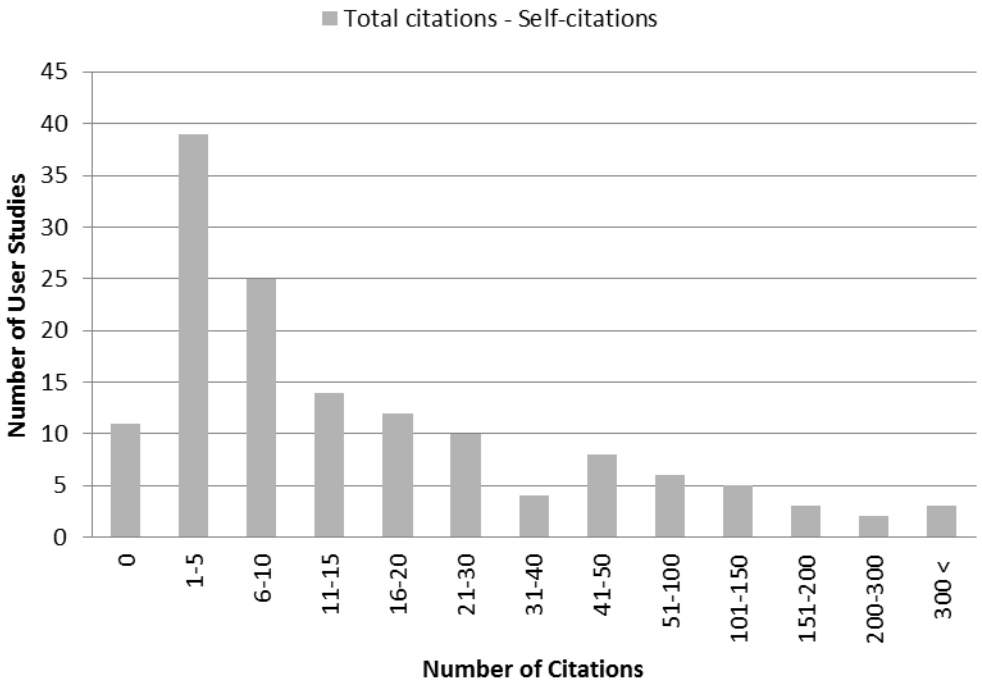


Figure 7. Distribution of the number of references of the user studies in other scholarly articles (excluding recent articles published in 2011 and 2012)

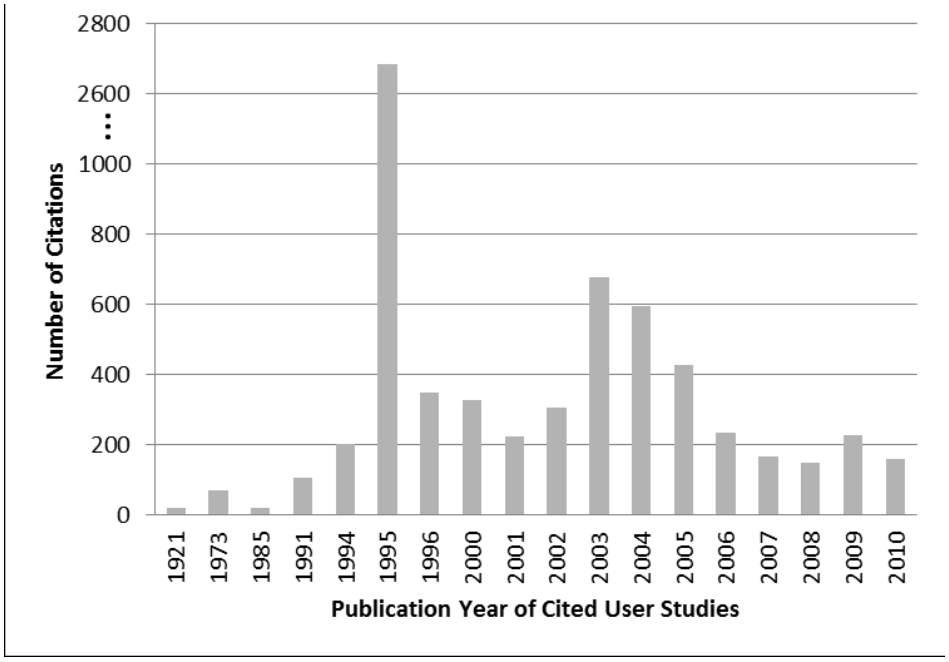


Figure 8. Distribution of the number of references of user studies by their publication dates

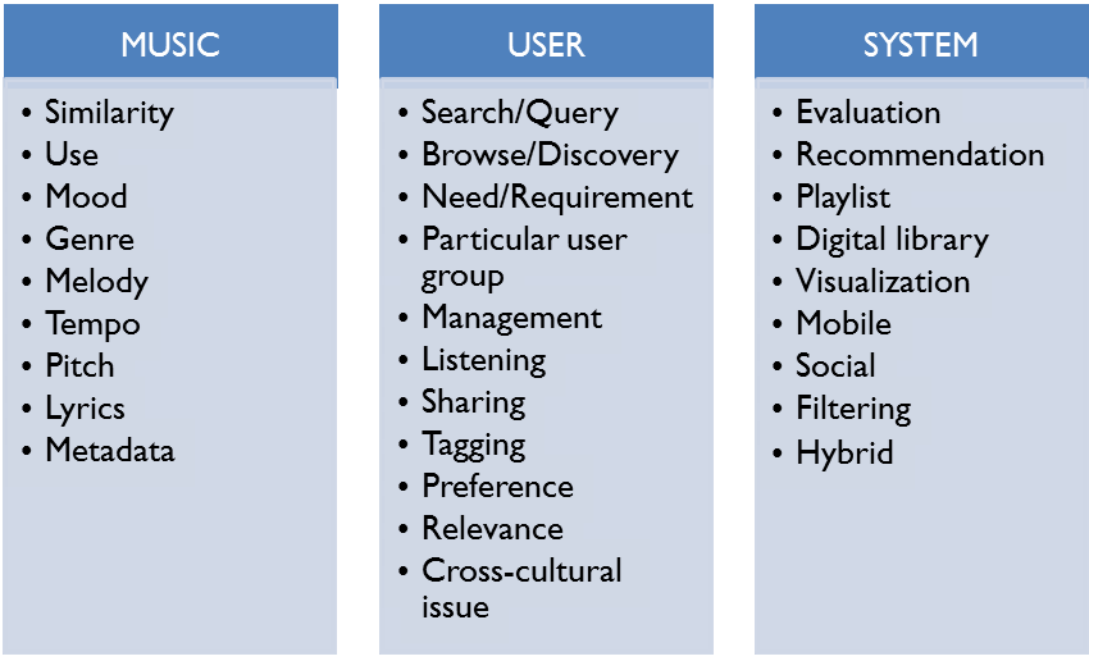


Figure 9. Taxonomy of topics in MIR user studies

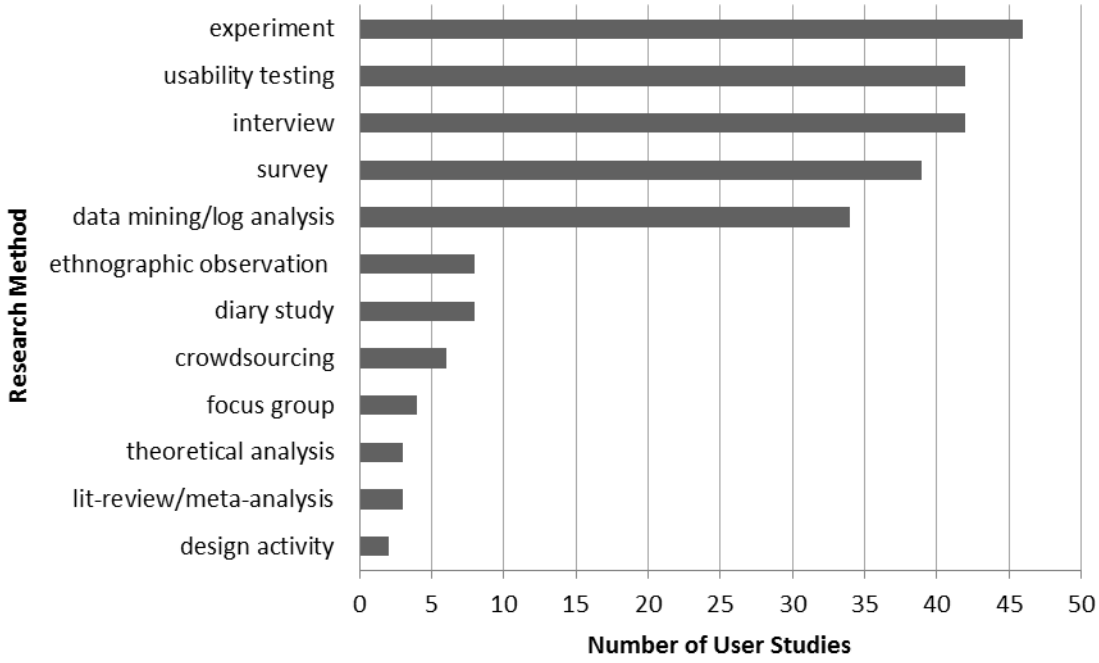


Figure 10. Research methods used in user studies

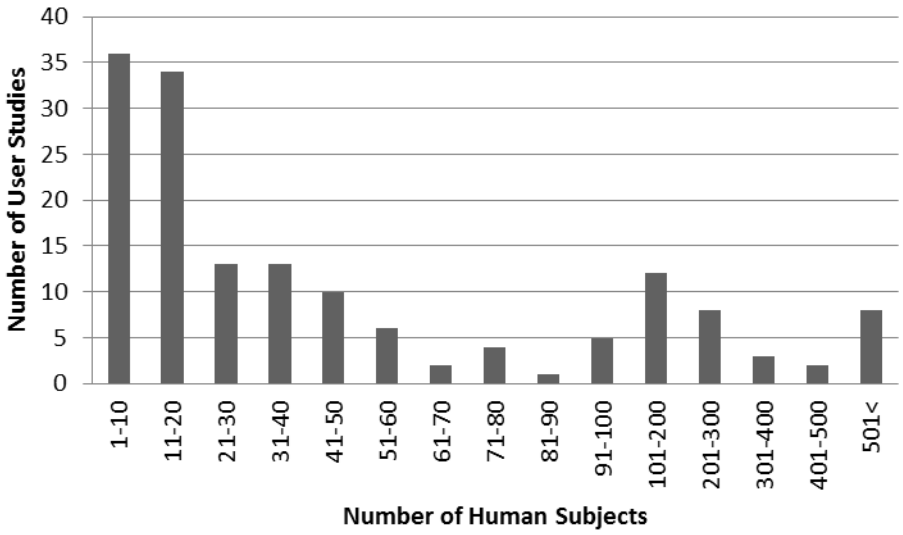


Figure 11. Number of subjects in user studies

Authors	Year	Title	Ref
Shardanand U, Maes P	1995	Social information filtering: algorithms for automating 'word of mouth'	2687
Rentfrow PJ, Gosling SD	2003	The Do Re Mi's of Everyday Life: The Structure and Personality Correlates of Music Preferences	406
McNab RJ, Smith LA, Witten IH, Henderson CL, Cunningham SJ	1996	Towards the digital music library: tune retrieval from acoustic input	347
North AC, Hargreaves DJ, O'Neill SA	2000	The importance of music to adolescents	257
Berenzweig A, Logan B, Ellis DPW, Whitman B	2004	A large-scale evaluation of acoustic and subjective music-similarity measures	235
Levitin DJ	1994	Absolute memory for musical pitch: evidence from the production of learned melodies	199
Bull M	2005	No dead air! The iPod and the Culture of Mobile Listening	165
Sloboda JA, O'Neill SA, Ivaldi A	2001	Functions of music in everyday life: An exploratory study using the Experience Sampling Method	163
North AC, Hargreaves DJ, Hargreaves JJ	2004	Uses of music in everyday life	142
Voida A, Grinter RE, Ducheneaut N, Edwards WK, Newman MW	2005	Listening in: practices surrounding iTunes music sharing	125
Ellis DPW, Whitman B, Berenzweig A, Lawrence S	2002	The quest for ground truth in musical artist similarity	116
Pauws S, Eggen B	2003	Realization and user evaluation of an automatic playlist generator	112
Boltz M, Schulkind M, Kantra S	1991	Effects of background music on the remembering of filmed event	105
Lee JH, Downie JS	2004	Survey of music information needs, uses, and seeking behaviours: preliminary findings	86
Cunningham SJ, Reeves N, Britland M	2003	An ethnographic study of music information seeking: implications for the design of a music digital library	78

Table 1. The top 15 most cited user studies (excluding self-citations)

Author	Number of user studies	Number of total citations	Number of co-authors
Cunningham, Sally Jo	14	682	16
Lee, Jin Ha	12	142	8
Downie, J. Stephen	11	253	7
Holm, Jukka	9	6	7
Inskip, Charles	7	18	4
Lehtiniemi, Arto	7	3	4
MacFarlane, Andy	7	20	4
Bainbridge, David	6	135	6
North, Adrian C.	6	446	7
Laplante, Audrey	5	30	1
Rafferty, Pauline	5	5	2

Table 2. The top 10 authors who published the highest number of user studies (excluding self-citations)

Unigram[a]	Count	Bigram[b]	Count
user	36	digit music	12
inform	23	music recommend	10
retriev	22	music collect	9
listen	20	everydai life	9
digit	19	music retriev	9
system	16	music library	7
evalu	16	music prefer	7
recommend	14	retriev system	7
librari	13	social network	5
collect	13	music genr	5
social	13	digit librari	5
		music listen	5

a. The most common term "music" (appeared 186 times) was removed.
b. Common phrases such as "music inform(ation)" (14 times) and "inform(ation) retrieval" (10 times) were removed.

Table 3. Top 10 unigrams and bigrams from the title words of the user studies

2003	2004	2005	2006	2007	2008	2009	2010	2011	2012
user	retriev	**listen**	**audio**	design	inform	user	user	user	listen
design	**digit**	**base**	**content**	evalu	retriev	**recommend**	collect	listen	digit
queri	**evalu**	**collect**	**seek**	**everydai**	**context**	collect	explor	explor	mood
retriev	**inform**	evalu	**playlist**	**find**	system	**prefer**	inform	recommend	role
sing	**need**	**new**	studi	**gener**	user	**social**	retriev	evalu	user
song	user	**system**	user	inform	**develop**	system	digit	interfac	design
studi				**mobil**	**librari**	**associ**	librari	mobil	**dj**
				role	**search**	base	**analysi**	playlist	experi
				user	design	evalu	**applic**	servic	**histori**
					digit	**explor**	associ	similar	**homeless**
					implic	genr	**criteria**	social	**influenc**
					interact	inform	**crowdsourc**	collect	inform
					mean	**interfac**	evalu	**comput**	librari
						investig	everydai	**discoveri**	life
						listen	**featur**	**engag**	**live**
						mood	genr	everydai	mean
						person	interfac	**experi**	**network**
						retriev	judgment	featur	**peopl**
						servic	life	interact	recommend
						tag	mobil	life	**share**
						vote	need	prefer	social
							person	retriev	song
							prefer	**select**	studi
							relev	**visual**	**young**
							search		
							similar		
							social		

Table 4. Top 15 ranked title terms over the past 10 years (w/ties); new terms are highlighted in bold.