TagFlip: Active Mobile Music Discovery with Social Tags

Mohsen Kamalzadeh[1], **Christoph Kralj**[2], **Torsten Möller**[2], **Michael Sedlmair**[2]
[1]Simon Fraser University, Burnaby, Canada, mkamalza@sfu.ca
[2]University of Vienna, Vienna, Austria, christoph.kralj@gmail.com,
torsten.moeller@univie.ac.at, michael.sedlmair@univie.ac.at

ABSTRACT

We report on the design and evaluation of TagFlip, a novel interface for active music discovery based on social tags of music. The tool, which was built for phone-sized screens, couples high user control on the recommended music with minimal interaction effort. Contrary to conventional recommenders, which only allow the specification of seed attributes and the subsequent like/dislike of songs, we put the users in the centre of the recommendation process. With a library of 100,000 songs, TagFlip describes each played song to the user through its most popular tags on Last.fm and allows the user to easily specify which of the tags should be considered for the next song, or the next stream of songs. In a lab user study where we compared it to Spotify's mobile application, TagFlip came out on top in both subjective user experience (control, transparency, and trust) and our objective measure of number of interactions per liked song. Our users found TagFlip to be an important complementary experience to that of Spotify, enabling more active and directed discovery sessions as opposed to the mostly passive experience that traditional recommenders offer.

Author Keywords

Music discovery; recommendation; user controlled; fine tuning; folksonomies; social tags; user-centred design; minimal effort; exploration; user interface; transparency

ACM Classification Keywords

H.5.2. Information Interfaces and Presentation (e.g. HCI): User Interfaces — Evaluation/methodology, graphical user interfaces, user-centered design

MOTIVATION

In the past decade, the amount of music we can have immediate access to has increased dramatically as music streaming services that provide access to millions of songs for a small subscription fee have risen in popularity. In the year 2015, most such services house more than 30 million songs. With such a broad range of choices, discovering new music and deciding what to listen to can become a burden. While this well established phenomenon [30] is not new to the age of music streaming, the effects of it have become more pronounced as the hesitation of having to pay for each individual track has diminished.

Recently, there has been a surge in both academic and commercial efforts to build interfaces and algorithms that can integrate these growing online libraries into our daily music consumption routines. Some solutions have focused on minimizing user interaction and relying on advanced recommendation algorithms. In such cases, the user's immediate role has been reduced to choosing a starting point (like a song, or a genre), and liking/disliking songs, with the rest being decided by other factors such as the user's preference profile, the context of listening, or various measures of content and user similarity. While these solutions excel in simplification, they suffer from issues such as lack of transparency[1], lack of user control, and pigeon-holing the users in their preference profiles [17, 37].

On the other end of the spectrum, elaborate interfaces have been developed that give the users control over various aspects of music retrieval, such as the parameters of a recommender algorithm. Although these interfaces have been shown to overcome some of the mentioned issues of recommender systems, they are generally complex and are designed with large screens and prolonged periods of user engagement in mind. This is in discrepancy with the typical situations in which we listen to music, such as when commuting and at work [15]. To fully exploit the potential of massive music libraries, it is crucial for novel methods of music discovery to naturally squeeze into these periods of everyday music listening. Commercial recommendation services like Spotify and Apple Music have identified this need and strive to provide simple interfaces to their vast libraries. This has, however, left them suffering from the usual issues that plague non-interactive recommender systems, as mentioned above.

With TagFlip we attempt to identify the sweet-spot between user control and interaction effort. Our goal is to put the user back in control of the music retrieval process while keeping the required effort minimal. Using a rich dataset of social tags from Last.fm, TagFlip morphs the conventional seed-based recommender that is well equipped for passive listening, into an interface that invites the user to explore and discover new music and musical styles by specifying tags—to perhaps break out of a comfort-zone and take a left-turn. At the same time, it minimizes the friction between listening and active discovery by diminishing both the mental load and physical interaction effort required from the user to initiate the discovery process, choose various styles of music to explore, and express or change his/her preferences.

[1] often referred to as the black-box issue with recommender systems

IUI 2016, March 7–10, 2016, Sonoma, CA, USA.
Copyright is held by the owner/author(s). Publication rights licensed to ACM.
ACM 978-1-4503-4137-0/16/03 ...$15.00.
http://dx.doi.org/10.1145/2856767.2856780

Identifying the necessity of adapting to the increasing mobility of music listening, TagFlip was designed with small screens (roughly five inches in diagonal) in mind. With minimal required effort, TagFlip can fit well into the periods of idle time that we typically fill by interacting with our phones. Sitting in the bus, exercising, taking a few moments of rest mid work, or sitting on the couch holding our phones and paying a bit of attention to the TV are all examples of these times.

We evaluated TagFlip in comparison to Spotify's mobile application in a lab study with 16 participants. Our users rated TagFlip on par with Spotify in usability and higher in three of our four recommendation constructs (*interface and interaction adequacy*, *control and transparency*, and *attitudes and behavioural intentions*). We also observed that TagFlip required significantly fewer screen touches for discovering a new liked song across all users. In our interviews, half of our users desired to have TagFlip on their own phones and almost all of them pointed out that it filled a crucial empty space in music recommendation, especially for specific and highly controlled discovery.

In summary, the main contributions of our work are twofold:

- Providing insight into the design and evaluation of tag based interactive music discovery tools, along with design considerations and topics for future research.
- Identifying the strengths of such systems and providing evidence that social tags can be effectively employed as direct means for user control in music recommendation.

BACKGROUND AND RELATED WORK

A review of state-of-the-art in music retrieval techniques and interfaces that are designed to exploit massive scale libraries reveals two dominant poles in terms of the level of user engagement. On one end of the spectrum are purely algorithmic approaches that minimize user involvement, while on the other end, we have elaborate and often complex interfaces that strive to put the user at the front and centre. These two extremes cater to the *satisficing* and *maximizing* behaviours that Schwartz et al. [31] explore. In the former case, the user is looking for something that is "good-enough", whereas in the latter, the user tries to maximize the degree to which the selected item adheres to his/her preferences at the moment.

Music consumption literature has confirmed how the listeners' behaviours fit this spectrum and how where the user ends up on it can depend both on the users' general enthusiasm and knowledge about music [11, 14, 16] and the listening context [4, 16]. Most current tools fall on the satisficing or maximizing extremes of the spectrum. Besides recommenders, the emerging notion of curated radio stations and playlists in services like Spotify and Songza is also very close to the satisficing end of the engagement range, with the added human touch in their creation. The role of the user is, of course, not neglected in algorithmic recommender systems. However, it lives as a personalized profile built based on various elicited or inferred preferences over longer periods of time, the inner workings of which are predominantly hidden from the user. Hence, while such systems excel at requiring absolute minimal interaction, when the user leaves the boundaries of pure satisficing, they fall short of providing any apparatus for increased user engagement and control. The popular like/dislike button that accompanies this type of recommendation is far from having any immediately visible feedback or a meaningful effect on the recommender algorithm. The user cannot specify what aspects of the music caused the like/dislike, and does not know what influence the action will have on future recommendations. Often times, prolonged use of these services accompanied with frequent like/dislikes leads to the user being enclosed in a bubble of his/her own computed taste profile with no way of knowing what went wrong or how to break out; an effect which is often called pigeon-holing in the recommendation literature [17]. On the maximizing end, we have the possibility of selecting specific musical entities like songs or albums, which bring about maximum user control on the music content at the expense of large interaction effort. Novel interfaces that fall in between are mostly designed for large screens and complex interaction and do not fit the current rapid shift to mobile music listening and discovery. As such, a rather unexploited space remains where minimal interaction and high user control meet, one that has also been identified by previous work [16].

The importance of transparency, control, and feedback during various stages of recommendation is well known [32], and has given birth to conversational and critiquing recommender systems [7, 8], which are dedicated to increasing user involvement in recommendation algorithms, through user feedback on attributes of recommended items. More recently, interfaces have been designed to explain recommendations to users [28, 41], or allow them to visually manipulate aspects and parameters of recommendation algorithms [5, 23, 40]. Faceted filtering and recommendation [46] is another more manual technique that allows the user to quickly combine various attributes in search. When it comes to music, such systems have for the most part not been designed for efficiency, simplicity, and mobile use and are predominantly based on a limited facet space [9, 45, 47], unlike TagFlip which houses more than 350 tags of various kinds.

An early effort toward increasing user control in recommendation of music was the MusicSun [26], which gave users lists of artists based on one of nine directions (*rays*) of similarity to a seed artist, with each ray representing a web-mined word. Another example is Tasteweights [5], which allowed the user to manipulate factors from his/her own taste profile, along with information from Wikipedia, Facebook friends' preferences, and experts from Twitter, to tune the recommendation list. While these and similar efforts are valuable steps in the right direction, they are first and foremost designed for prolonged sessions of use or come with interfaces that are often too elaborate, fitting larger than mobile screens.

A body of research has focused on visualizing large music libraries to simplify their exploration. This generally involves depicting the musical entities, such as artists or songs, on a 2D or 3D map with the assistance of dimensionality reduction techniques, and allowing the user to traverse or zoom the map. The input to these techniques can be any type of descriptor, like features extracted from the audio itself, or metadata. Some pioneering studies on this front include MARSYAS3D by Tzanetakis and Cook [38], Islands of Music by Pampalk et al. [27], and the artist map by van Gulik et al. [39]. In general, these types of interfaces either do not offer sufficient user control, or suffer from issues similar to that of controllable recommenders discussed above, namely, a need for large displays and more indulgent exploration, rather than efficient and minimal interaction.

Identifying the need for simplicity coupled with higher possibility of user control than what pure satisficing interfaces facilitate, Baur et al. [1] introduced the *Rush* technique, which gave the user a choice between multiple recommended items, each of which could sway the algorithm in a different direction than others. Unlike TagFlip, the interface was built for creating homogeneous playlists from songs already known by the user, rather than discovery and song by song recommendation. A later version of *Rush* [2] added the possibility of controlling the similarity to operate based on artist, genre, tempo, or songs form the same artist. In comparison, TagFlip provides finer grained control over music selection by utilizing social tags, which cover a much broader range of music descriptions. Furthermore, the user can easily combine various tag filters using TagFlip, but *Rush 2* does not allow combining similarity axes.

A number of tools have previously used social tags as basis for interactive recommendation. Vig et al. [42] used tags of movies in a critiquing-based recommender, allowing users to decrease or increase the weight of each tag of a recommended movie to get the next one. While their approach bears similarities to ours in using a rich set of social tags, the interface is not built with music in mind, and does not fit the dynamic and song-to-song nature of music listening. Wang et al. [44] built an interface for querying music with multiple weighted tags. With this tool the user can click and drag on a tag in a tag cloud to increase or decrease its importance in recommendation. The study is, however, mostly algorithm oriented and explores the retrieval techniques rather than user experience. In a similar fashion, the Music Explaura system [12] used interactive tag clouds (*textual auras*) as basis for recommending artists. Users of this system experienced a steep learning curve. Although they expressed interest and surprise at the concept once explained to them, most did not immediately realize the meaning of the tag clouds and the fact that they could manipulate them. Meerkat [25] also used tags as means for personalizing radio stations. Unlike us though, the authors focused only on the functionality, finding that people liked having this level of control on their music. However, they left out design completely and the interface was not tested for usability. Furthermore, none of the above tools were designed for mobile devices.

With TagFlip, we went through an interative user centred design process to identify and address the key design issues involved with bringing tags into interactive discovery and recommendation. In the next section, we explain the various design decisions made through-out this process and provide implications for design of future tag based tools.

TAGFLIP

In this section we first discuss how we selected and processed data for TagFlip. Then, we explain our interaction paradigm and design requirements for the interface and report on what we learned throughout the design process.

Data and platform

Among the myriad types of data that can be utilized as the basis for music recommendation, classification, and retrieval, few are easily understandable for the average non-technical music listener. While recommendations based on collaborative filtering or audio content can be accurate according to algorithmic precision measures, the data underlying such algorithms is not translatable to tangible and easy to interpret attributes of music, as such, it is not directly user controllable.

On the contrary, the massive tag spaces formed in social music tagging platforms such as Last.fm can be rich sources of semantic music attributes that are understandable to the average user. These spaces (often called folksonomies) can play a key role in bridging the semantic gap between the users' description of music and how recommenders work. The terms present in the Last.fm folksonomy range from arguably every music genre and sub-genre imaginable, to moods, activities, and niche musical terms popular in smaller communities. Each song/tag annotation also has a score attached to it, which represents the percentage of taggers of the song who used the term to annotate it [36].

On the down side, with all this information come several issues, a variety of which and possible approaches to addressing them have been discussed by Lamere [18]. These include issues such as the cold-start problem (unpopular music gets very few tags), synonymy (multiple tags having the same meaning), polysemy (tags having multiple different meanings), and noise (spelling errors and terms with no meaning in the music domain). To alleviate some of these, our first step in employing Last.fm tags in our tool was a robust preprocessing phase, in which we utilized the available information on the popularity of songs (listener and play count), the scores of tags, and language processing techniques to remove meaningless or subjective tags, fix spelling errors, identify meaningful compound terms, extract usable information out of unusable tags, and remove subjective or vague terms. This process is explained in detail in supplementary materials. In the end, we had a cleaned set of 358 tags.

TagFlip was built for Android, and our music library contained 100,000 songs and 1.3 million song/tag associations. The music was a subset of the Million Song Dataset [3], which covers a broad range of contemporary music, and the audio was played from Spotify, using its Android SDK [34].

Interaction paradigm

The core user interaction in TagFlip consists of a repeated two phase exchange between the system and the user. The beginning of playback resembles a conventional recommender or search system. The user can either start from a specific song, or describe the desired music using a combination of tags. Once a song is played, TagFlip displays its top tags (Figure

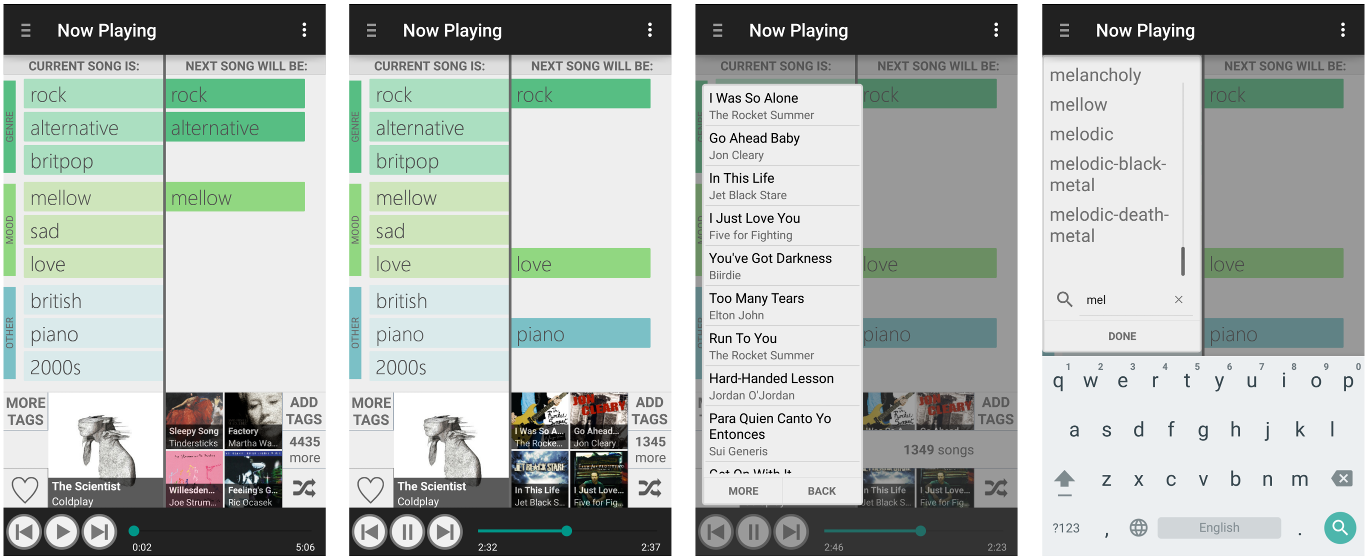


Figure 1. Screenshots of TagFlip. (a) What is displayed when a song is played from library. (b) After the user has modified the tags for the next song. (c) Viewing a list of upcoming songs. (d) Adding tags from a full list.

1(a)), and indicates which ones are being considered for retrieving future songs. The tags are grouped into genre, mood, and "other" categories. By default, three tags are "pinned", two genre tags and one mood tag. These define the future direction of music playback by constraining the library to songs that match all of them. A playlist based on this direction is built and set for the user who does not desire further engagement. This constitutes one phase of the exchange. The album arts of the next four songs are displayed in the bottom right of the screen, the number of songs that match the selected criteria is shown, and the complete list of songs can be accessed by tapping this number (Figure 1(c)). A shuffle button below the number of songs does the task of randomizing the order of upcoming songs.

The second phase involves the user modifying the set of tags used in finding future songs by simply tapping on any of the tags to pin or unpin it (Figure 1(b)). Each tap immediately updates the set of planned songs and this is made visible to the user through updating the number of matching songs and the next four album arts. Once the next song is played, the interface is updated to reflect its tags while keeping the previously pinned tags intact (which are by design also present in the new song), and thus, the exchange continues. In summary, the key factors that differentiate this core interaction paradigm in TagFlip from earlier efforts are:

1. High impact interaction instances: This is crucial for achieving high control coupled with minimal interaction effort. Each pinning or unpinning of a tag can greatly influence the set of future songs. In addition, as most songs have several diverse tags associated with them, one pin/unpin can lead the user to music that is similar to the previous song in some aspects and vastly different in others. This way, the user covers a larger span of the library than what would be possible with a conventional seed-based recommender system, without completely changing seeds or switching playlists.

2. Fine tuning: Another key differentiator between TagFlip and previous tools is how easily it enables fine tuning based on tags of the currently playing song. The lack of such functionality in current music discovery tools and a need for it has been stressed out by Kamalzadeh et al. [16], who call it *adjusting control*. Many of the participants in our final user study were particularly excited about this concept.

3. Low interaction effort: Previous studies have shown that users are interested in meaningfully altering aspects such as the mood of their music listening session with efforts as low as what skipping a song requires [16]. We used this as a guideline to what each tag pin/unpin should need. The only overhead of a pin/unpin compared to a skip is finding the desired tag on the screen. This leads us to the next factor.

4. Low mental load: To minimize the mental effort for selecting which tags to pin/unpin, TagFlip shows a summarized set of each song's tags categorized into "genre," "mood," and "other" terms. This gives the user an overview of the possibilities and obviates the need for extensive thinking. If a user desires, (s)he can view all other tags of a song using the "more tags" button and pin them with the same tapping action.

5. Scalability of control: Following the satisficing/maximizing spectrum and guidelines provided by earlier studies [4, 16], a key design requirement for TagFlip was the possibility to organically scale from coarse to fine-grained user control in conjunction with low or high user engagement. By allowing the users to pin as many tags as they wish, the system supports a range of control on the retrieved songs. With few pinned tags the system resembles a conventional recommender or radio station, but pinning tags can rapidly increase user control. If further control is needed, the user can quickly bring up a list of all tags in the system as a pop-up on the main screen (Figure 1(d)), using the "add tags" button, and quickly search and add tags from it. This addition is directly visible beneath the pop-up dialog.

6. Transparency: This is one of the key pillars of building trust and increasing user satisfaction with recommendation systems [37], the lack of which leads to the black-box issue. The two phase exchange in TagFlip is by nature explanatory and transparent. The user can clearly see the effect of each pin/unpin in the number of matching songs and the next four album arts. Furthermore, the fact that the number keeps decreasing as more tags are pinned informs the user that the songs will match all the pinned tags, rather than just any of them. As was proven in our lab study, this mechanism made it clear to users how the system worked and resulted in high user perceived transparency.

7. Small screen: Staying within the confines of smartphone screens underlined all steps of TagFlip's design. Apart from the necessity to acknowledge the rapid migration of music consumption to mobile devices, adhering to such limits would simplify redesigning the interface for a desktop environment while still conforming to the minimal interaction effort requirements. The small screen size helped us narrow down the absolutely critical user tasks to support and forced us to think in terms of "what should be left out" instead of "what should be added". Several usability tests during various stages of design (from low to high fidelity prototypes, to an actual application) informed these decisions.

Design process and decisions

TagFlip started from the idea of the user being in control of high-level aspects of music. In its path from paper prototype to working application, which took 18 months, several design possibilities were explored and tested. Three stages of formative usability tests with a total of 10 different participants were performed through-out this process. Out of nine initial designs, two made it to the medium fidelity prototyping stage which was done in a desktop Java environment, and one to Android. We now elaborate on a few key findings of this process and the choices that were made according to them.

Individual tag blocks or a left/right flow:

Coming out of paper prototypes, we had two thoroughly different competing designs for TagFlip (available in supplementary materials). In one, each tag was housed in a separate square block, and there was no high level separation between the current and next songs in the interface. Instead, each block displayed the strength of each tag in the currently playing song and provided a slider for choosing the strength for the same tag in the next song. This design was based on the faders of a mixing board. Our usability tests showed this to be a confusing system image and lacking organization, with many users not noticing the current and next separation in each block. This lead us to choosing the current presentation, where a flow from left to right implicates a shift from current to next, resembling how we read and write in English. We also found it was important to clearly separate the two sides of the interface to further amplify the current/next concept. This is what lead to the addition of the vertical line in-between, and every UI element on each side only pertaining to its corresponding time stamp; current or next (except for the play bar in the bottom).

Strength of tags:

One of the earliest ingredients of TagFlip was the ability for the user to specify how strong the association between the retrieved music and the pinned tags should be, as an added level of user control. This is similar to what the interactive tag cloud systems discussed earlier provided [12, 44]. The earliest prototypes explored ways of supporting such a task by the use of sliders, interactive bars, knobs, and text size. However, heuristic evaluations and usability tests with users revealed that such a functionality could lead to a confusion between strength of the tag and how much they cared about it. Also, for some users, the tags had a binary meaning; they either belonged with a song or did not. Our tests showed that most users did not care about having this functionality at all. This fact, plus the added complexity that such a feature would incur on the interface, lead us to removing it from the later versions of TagFlip.

Communicating target set sizes:

To provide full transparency, TagFlip employs a strict retrieval policy, where all the target songs should include all the user requested tags. Therefore, adding too many tags or certain unpopular combinations can lead to an empty set of songs. This could also happen if all the songs in the set have been played without the user changing preferences. In such a scenario, if a song has to be played (through tapping the next button or the current song ending), TagFlip automatically removes the least popular tag to retrieve music. In our design process, we considered adding visual encodings that would either prevent the user from pinning tags that would lead to an empty set, or help the user easily rectify the problem if such a thing were to happen. We used horizontal bars placed on the right end of the screen in front of each tag. The size of each bar reflected the size of the resulting set of songs, if the corresponding tag was to be added. Once a tag was added, the bar changed colour, and its size indicated how much one would expand the target set if the tags was removed. However, our usability tests proved that this encoding was often confused with the above *tag strength* concept. Hence, the encoding was completely removed. In the final version, the interface would just inform the user that no songs were found, and would suggest the removal of the least popular tag.

Categorizing tags into genre, mood, and other:

Previous studies have shown genre and mood to be two of the top attributes of music for users in selecting/looking for music or managing their libraries [15, 35, 43]. Therefore, we found it sensible to categorize tags into these two, as a way of reducing cognitive load. The "other" category was added to house terms that could not be classified as genre or mood. Context of listening has also been reported to play an important role in music selection, but words relating to context were not popular in our dataset, making it pointless to place them in a separate category. We decided that the three top terms from each category could sufficiently describe a song. Categorization was done based on lists of genres and moods which we built by combining and cleaning various lists from the web. In cases were these lists intersected (e.g. "romantic") we prioritized the genre list. It is noteworthy that since user input is never dependant on this categorization, the accuracy of these lists is not of importance; they are used solely to organize the presentation.

Limiting the main screen to nine tags:
Early prototypes of TagFlip presented all tags of the currently playing song in the main screen and made each category scrollable. This created visibility issues with part of the pinned tags being hidden if scrolled out of the screen. To mitigate this, we limited the number of tags shown by default for each song to nine. Where available, these were equally distributed into the three above explained categories. If a category contained less than three tags for a certain song, we filled the screen by including more than three tags from other categories. The user had the possibility to view the rest of a song's tag by clicking the "more tags" button in Figure 1.

EVALUATION

To evaluate TagFlip, we compared it to Spotify's mobile app, which has a library of more than 30 million songs, in a lab study designed around comparing the applications in their music discovery capabilities. The setup was refined through three pilot tests, and 16 participants (8 female, median age = 26) were recruited for the main study. Nine of these had a background in computer science, and nine had a Bachelor's or higher degree. We chose Spotify as a point of comparison as it is one of the most popular music streaming/recommendation services, and most of our participants either used it regularly or had experience with it. We decided against comparing TagFlip to an in-house made conventional recommender (with seed attributes and like/dislikes) because it would end up as just a less powerful version of TagFlip with no tag based navigation, and would presumably give us unrealistically positive results. With this test, we intended to investigate whether there is room for an interface like TagFlip in our user's daily music consumption, and to that end, the sensible route was to compare it to the state-of-the-art in commercial systems. Besides similar music recommendation (radio stations based on songs or genres) and community made playlists, Spotify also houses features such as top charts, artist pages and albums, and similar artists. All of these were open to participants.

We employed a within subjects design with a mixed methods approach [19] which consisted of questionnaires, interviews, videos of each session, and usage logs (with TagFlip only). The participant first filled a questionnaire covering basic music listening habits such as average hours of listening per day and use of streaming/recommendation services, and demographics. Then, the participant used each interface for 10 minutes (balanced order between participants). The task was to find new songs that (s)he liked and had not heard before (or had heard a long time ago and forgotten about), and to save the found songs by tapping the heart button in TagFlip, or adding them to a playlist in Spotify. Before the 10 minute main task, each user was also given five minutes with each interface in order to get familiar with it. During this time, the participant could ask questions about the interface from the experiment conductor. After finishing the 10 minute task with each interface, the participant filled two questionnaires according to his/her experience with it; one for general usability, and one for recommendation aspects. For the former, we used the SUS questionnaire [6], and for the latter, we built a 22 item questionnaire with 5 point Likert scale answers, based on the ResQue framework [29]. After both tasks were performed, a semi-structured interview was conducted, which mainly revolved around how the participants usually discovered new music, and how they would compare the two interfaces. The study was done on an HTC One M7 phone (4.7 inch screen) which was connected to loud speakers. For each participant, the study took roughly one hour, after which the participant was compensated 10 Euros.

Following the ResQue framework [29], we categorized our questions on recommendation aspects into four main constructs: **(1)** Quality of recommendations, **(2)** Interface and interaction adequacy, **(3)** Control and transparency, and **(4)** Attitudes and behavioural intentions. Based on these constructs, number of user interactions, and the liked songs, we formulated 7 hypotheses categorized into three higher level ones and four sub-hypotheses about each construct in our recommendation questionnaire. These seven were designed to question both the subjective and objective user experiences with TagFlip and Spotify:

- ***H1:*** *The overall user rating for recommendation aspects of TagFlip will be higher than Spotify (aggregate score based on all 22 questions)*
- ***H1.1 to H1.4:*** *TagFlip will be rated higher than Spotify in all the four constructs of recommendation.*
  - ***H1.1*** *Quality of recommendations*
  - ***H1.2*** *Interface and interaction adequacy*
  - ***H1.3*** *Control and Transparency*
  - ***H1.4*** *Attitudes and behavioural intentions*
- ***H2:*** *The number of interactions (screen touches) per liked song will be smaller for TagFlip.*
- ***H3:****The number of songs liked will be larger for TagFlip.*

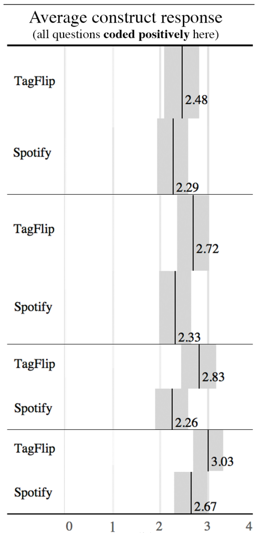
H1 and its four sub-hypotheses gauge the subjective reaction from our participants, As discussed earlier, the key differentiators of TagFlip relate to precise user control and high transparency, coupled with low interaction effort. These factors align with the *control and transparency* and *interface and interaction adequacy* constructs from our questionnaire (H1.2 and H1.3). We predicted that better performance in these two categories would also lead to an overall better experience with TagFlip (H1), and better user rating in the other two constructs (H1.1 and H1.4). H2 was intended for helping us objectively test whether TagFlip actually required small interaction effort for exerting high control and reaching desired music. Finally, H3 was born out of the assumption that a better experience with TagFlip would lead to more liked songs.

Questionnaire results

Eleven participants said they used online streaming services on a regular basis, and the most popular service was Spotify. The medians for size of personal music collection, active listening hours (focused), and passive listening hours (during

Construct	Question Type	Question	TagFlip	Spotify
Quality of recommendations	Positive	The recommended songs matched what I was looking for	3	3
		The recommendations I received better fit my interests than what I may receive from a friend	**2**	1.5
		I enjoyed the recommended songs	3	3
		The recommender system helped me discover new music that I like	3	3
	Negative	A recommendation from my friends better suits my interests than the recommendations from this system	2	2
		The recommended songs were not enjoyable for me	**0.5**	1
		I could not find new music that I liked using the recommender	**0**	1
Interface and interaction adequacy	Positive	The app made it clear why the songs were recommended to me	**3**	1.5
		The app provided an adequate way for me to express my preferences	**3**	2.5
		I easily found the recommended songs in the interface	3	3
		I found it easy to find songs that I liked using the recommender system	3	3
	Negative	It was difficult to tell the app what I was looking for	**0.5**	2
		Looking for a list of recommended songs required too much effort	2	**1**
		It required too much effort to find what I liked using the recommender system	1	1
Control and Transparency	Positive	I felt in control of telling the app what music I wanted to find	3	3
		I understood why the songs where recommended to me	**3**	2
	Negative	It was difficult to understand why some songs were recommended	1	1
		I did not feel in control of specifying and changing my preferences	**0**	1
Attitudes and behavioural intentions	Positive	Overall, I am satisfied with the recommender	3	3
		I think I would like most of the songs that I find with this app if I use it again	3	3
	Negative	The recommender can not be trusted	**0.5**	1
		The recommended songs confused me	**0.5**	1

(a)



(b)

Figure 2. Part (a) shows median responses for all the questions regarding recommendation aspects, coded from 0 to 4, with 4 being the best response for positive questions and 0 being the best for negative ones. In part (b), average scores for all constructs along with 95% confidence intervals are shown. Here, negative questions were reverse coded to contribute positively.

other activities) were 5895.5 songs, “30 minutes to 1 hour”[2], and “2 - 4 hours”[3] respectively.

On the usability side, both interfaces scored similarly on the SUS questionnaire[4] (TagFlip: $M = 75.63, \sigma = 10.97$; Spotify: $M = 72.66, \sigma = 18.67$) with no significant difference in a paired sample t-test; $t(15) = 0.54, p = 0.60$. For some participants, the main screen in TagFlip was not immediately clear at first glance. However, they quickly realized the interaction concept and how the songs were retrieved once they started tapping on the tags and observing the next four songs and the size of the target set being updated and decreasing with each added tag, in the bottom right of the screen.

Figure 2 shows the median response to all the questions regarding recommendation aspects for both interfaces, classified into the four above mentioned constructs, and further grouped into positive and negative. Responses were coded into scores from 0 to 4. In positive questions, 4 is the best response, and in negative questions 0. For each question, the better median value is highlighted. As shown in the table, TagFlip tops Spotify in 11 of the 22 questions, while it's beaten in one. Using reverse codes for negative questions, the aggregate score over all 22 questions turned out to be significantly higher for TagFlip (TagFlip: $M = 3.03, \sigma = 0.51$; Spotify: $M = 2.51, \sigma = 0.66$; $t(15) = 3.12, p < 0.01, d = 0.78$). This **confirms H1**. Response histograms for all 22 questions are provided in supplementary materials.

Looking at individual constructs, TagFlip scored significantly higher in three of the four. These were *interface and interaction adequacy* ($t(15) = 2.26, p = 0.04, d = 0.56$), *control and transparency* ($t(15) = 3.86, p < 0.01, d = 0.97$), and *attitudes and behavioural intentions* ($t(15) = 2.54, p = 0.02, d = 0.63$). These numbers **support H1.2, H1.3, and H1.4** as well. In the *quality of recommendations* construct, although TagFlip scored higher, the difference was not found to be significant ($t(15) = 1.88, p = 0.08$), **leaving H1.1 unconfirmed**.

Objective measures

To test H2 and H3 we analysed the videos captured from the participants' interactions with both interfaces. We counted each touch of the screen (including a non-broken scroll) as one interaction. Since the act of adding a song to a playlist in Spotify, which was what the participants were instructed to do if they liked a song, required three taps (compared to one tap in TagFlip) we counted each of these three tap sequences as one interaction in Spotify. Each instance of typing in a search box anywhere in the interfaces was also counted as one interaction.

The number of interactions with TagFlip were smaller for all but one participant, and this was found to be a significant effect; TagFlip: $M = 144.56, \sigma = 37.00$; Spotify: $M = 209.56, \sigma = 54.64$; $t(15) = -5.85, p < 0.01, d = -1.48$. A look at our videos made it clear why such a large difference existed. With Spotify, participants spent a lot of time switching between various charts, playlists, and radio stations; actions that need several taps. In contrast, in TagFlip most of the interaction happened in the main screen with the tags or the list of upcoming songs.

Beyond that, we were specifically interested in how effective the participants' interactions were. While it is generally difficult to objectively measure such a concept between two fundamentally different interfaces, we chose to measure the number of interactions per liked song as a way to approximate effectiveness. This value was found to be significantly smaller for TagFlip ($M = 25.61, \sigma = 15.07$) than Spotify ($M = 34.58, \sigma = 18.47$); $t(15) = -2.05, p = 0.04$. As such,

[2] Asked in ranges: <15m, 15m -30m, 30m - 1h, 1h - 2h, > 2h
[3] Asked in ranges: < 1h, 1h - 2h, 2h - 4h, 4h - 6h, > 6h
[4] SUS gives a score between 0-100

our data supports H2 as well. Looking just at the number of liked songs, we saw very close numbers for the two interfaces for most users (TagFlip: $M = 7.00, \sigma = 3.10$ Spotify: $M = 7.06, \sigma = 3.09$), with no significant difference found in a paired sample t-test ($t(15) = -0.79, p = 0.938$). As such, our **results do not support H3**.

Another measure to look at is the number of tags pinned when recommendations were played. With Spotify, the user has only one point of control, which is realized through selecting a seed song for recommendation or a pre-compiled playlist. With TagFlip, participants had an average of 2.27 tags pinned through-out their 10 minute exploration, which indicates a level of desired control higher than what one tag can achieve. A comparison between recommendations that lead to a like by the user and those that did not, reveals a significant difference between the number of pinned tags in the two cases; Liked: $M = 2.46, \sigma = 0.74$; not liked: $M = 2.24, \sigma = 0.75$; $t(15) = 2.382, p = 0.03, d = 0.60$. This can suggest that either the songs chosen with more specific constraints turned out to be of higher quality for the user, or that participants were more likely to think they liked a song if they applied more control toward retrieving it.

Interview results

We manually coded the interviews to identify prevalent concepts. The most prominent comment about TagFlip (mentioned by all participants) was the level of control it provided to the user. Participants appreciated the fact that it was easy to specify exactly what type of music they liked to hear as opposed to the more unpredictable experience with Spotify. As one user put it in simple terms *"You made tags useful"*. Another said jokingly, *"Sometimes you have your own ideas about what your music should be. It is not always what the powers that be think about it."* Another participant said *"in Spotify, you get a radio, like it or not!"*. Some of the participants who used TagFlip second actually complained about the fact that they could not combine criteria in Spotify, and some even asked whether they had missed the feature in its interface. One participant said that with Spotify, she kept trying to reach a *"pleasant stream"* but was not able to. With TagFlip on the other hand, she could easily reach that state; she could choose a number of tags and then expect to like most of the songs and *"prevent radical changes"*. Seven participants actually mentioned that they did not like Spotify's recommendations or playlists because of too little control. On the other hand, eight participants also liked the fact that Spotify could sometimes require a smaller effort and that playlists were human made. For instance, one user mentioned the fact that you sometimes need some *"up and down"* and a *"good mix"*, which the community made playlists can provide.

Another popular concept among the gathered responses was what we call *fine tuning*. Eleven participants specifically liked how TagFlip let them tune their experience based on the tags of each played song. One of our users called this *"local control"*, complaining that no other tool he had used supported it. *"It helped me narrow down my mood"* was how he summarized this experience. We observed this phenomenon first hand as well. Participants would often see a tag and react to it with amusement or add it to the next song. *"hah, 'sexy', why not?!"* and *"'groovy', exactly!"* were examples of this. A similar comment noted how easily one could get diverse styles of music from TagFlip. As one participant put it *"I could easily go from blues to funk to rock to 70's, so I got a lot more"*. Other salient themes in interview responses related to the size of the next four album arts being too small (6 users), and appreciating the system's transparency (5 users).

Based on how excited the participants were about TagFlip, and whether they explicitly asked to have the app on their own phones (without us mentioning it) we classified them into two groups (ENTH: enthusiastic, REST: rest of the users). Seven participants belonged to ENTH. Out of the nine members of REST, another seven were still open to using TagFlip for active discovery, but were not as excited as ENTH members. Looking at the interview transcripts, we found that all members of ENTH mentioned the *fine tuning* aspect of TagFlip, while only four of the second group did so. Using Fisher's exact test, this was found to be a significant effect ($p = 0.03$). A similar association was seen for people who mentioned that they did not like Spotify's recommendations (5 from ENTH and only 2 from REST), however, this was not found to be significant ($p = 0.13$). We also observed a significant difference between the groups in terms of the differences between the number of liked songs with the two interfaces (ENTH: mean of differences between number of liked songs in TagFlip/Spotify = 1.71, $\sigma = 2.70$; REST: $M = -1.4, \sigma = 2.87$; $t(13.44) = -2.26, p = 0.04, d = -1.23$). This indicates that our enthusiastic group liked more songs in TagFlip than Spotify, compared to the rest of the participants. In addition, the earlier-mentioned significant difference between the number of interactions per liked song was created by ENTH users (ENTH: mean difference in interactions per liked song = $-19.32, \sigma = 14.15$; REST: M = $-1.24, \sigma = 14.95$).

DISCUSSION

Is there room for TagFlip?

While there is plenty of existing commercial and academic work on music discovery, most of it has focused on algorithm perfection rather than actively engaging the user. Hence, with this paper, our main goal was to foray into the less explored space of user-controlled music recommendation on mobile devices, and test the possibility of utilizing social tags to accommodate such control.

Out of our five hypotheses on subjective experience, four were supported in our results. These included the aggregate user feedback for recommendation aspects (H1), interface and interaction adequacy (H1.2), control and transparency (H1.3), and attitudes and behavioural intentions (H1.4). A careful look at individual questions reveals the main driving force behind these to be the way TagFlip provided means for users to easily manipulate and fine-tune their recommendations, and how it exposed its logic of operation to them (H1.2). This lead to a high level of perceived control and transparency (H1.3). Although the improvement in the overall quality of recommendations was not found to be significant (H1.1), the participants' trust toward TagFlip greatly benefited from its transparency and high control, leading to a significant difference in the fourth recommendation construct as well (H1.4). This is an interesting phenomenon, suggesting that the users' trust does not necessarily depend on the quality of the recommended items, but also on how much agency they had in the process. A similar effect was reported by McNee et al. [22], who found that higher control leads to higher loyalty, even despite more user effort and comparable recommendation accuracy.

Analysis of our interviews revealed a strong relation between mentioning the *fine tuning* aspect of TagFlip and being in the enthusiastic group of participants who asked to have TagFlip on their own phones as soon as possible. This concept has been largely absent from music recommenders. With current tools, to achieve similar results, users would have to identify the keywords they are interested in, and then manually type them in a search box and hope for the best. Besides minimizing the effort in both high-level and fine-tuning control, our way of presenting music can also lead to spontaneous explorations when it shows an intriguing tag. Our users often chose tags out of amusement, wonder, or sheer curiosity in looking at the number of conforming songs for various tag combinations. As such, TagFlip also enables the user to easily change course significantly with some tags, while keeping other tags pinned, to simultaneously achieve overall similarity.

Moving on to objective measures, we observed considerably different behaviours in user interaction between the interfaces. As mentioned before, most of our users' interactions with Spotify were directed at navigating the various UI elements, such as scrolling through lists of songs and playlists or switching between them. These actions incurred a large number of screen touches, significantly more than with TagFlip, where most interactions happened in the main screen, pinning/unpinning tags or scrolling the list of upcoming songs. Data from these observations supported our hypothesis on interaction effort (H2), which indicates that TagFlip was successful in keeping interaction effort minimal; one of its core design requirements. That said, the number of liked songs with the two apps was surprisingly close, with similarly close and relatively small standard deviations. This could hint at an unknown factor playing a role here. Perhaps this measure is not appropriate for gauging overall performance with the apps, as it might be rooted in an unconscious tendency to like an equal number of songs with both. Nevertheless, we believe that the above difference in interactions per liked song can still be meaningful and indicate a smaller effort required by TagFlip for finding a comparable number of new songs.

Our results suggest that music discovery based on social tags can be a viable solution to increasing user control in music recommendation. More importantly, our iterative design process gave us valuable insight into the users' expectations, reactions, and mental models regarding tag-based discovery. Our principal challenge in designing TagFlip was figuring out the appropriate way of laying out all the required information for the user within the confines of a mobile device, without causing cognitive overload and requiring intricate interaction. Our ability to design this interface with high usability and user satisfaction compared to one of the most popular commercial tools suggests that there is room for expanding tag-based interfaces out of the academic space and integrating them into consumer-facing services, in order to make user interaction with massive music libraries more efficient, directed, and transparent. We will now discuss some of the key findings of our design process.

Design considerations and remaining questions

TagFlip went through major changes in its prototyping stages. Some of our designs, beginning from paper prototypes and all the way to the current iteration, are included in supplementary material. Earlier, we discussed a few of our design choices in building TagFlip. Some of our key findings in the final evaluation were the following:

Full tag list presentation:
TagFlip uses an alphabetically ordered list of tags with the option to search, if the user desires to start listening by specifying tags rather than from a song. Some users found the mental load associated with this task to be too heavy. An interesting question would be how the list of all tags can be presented in a way that reduces the friction of thinking about what to choose. One solution could be a sorting of the list based on the users' listening histories.

Control on tag strength might improve satisfaction:
We discussed how tag strength was left out of the final interface due to the confusion between it and how much users cared about a tag or how popular it was. In our lab study, non of our participants expressed a need for such a feature. However, perhaps its addition can increase user satisfaction by having higher quality recommendations. TagFlip prioritizes songs that have all the requested tags over those that have some but with higher strength. While our participants were mostly content with the recommended material, in a couple of instances they complained that a tag did not belong with a song. Hence, future work can investigate the importance of such a feature and whether it can enhance user experience.

Communicating target set sizes; probably not required:
The tested design in TagFlip (informing the user that no exact match was found and encouraging the removal of a tag) proved to be sufficient in dealing with empty target sets. In the few times that this happened in our lab study, users quickly reacted to the message by removing the tags they cared least about.

Categorizing tags helps:
A third of our participants liked the categorization into mood/genre/other, one felt it was not needed, and the rest did not have specific opinions about it. As such, we did not find enough evidence to suggest that alternate designs could perform better in giving the user a high level idea of the type of music that is played.

Limiting the main screen vs. full scrolling:
Although showing only nine tags on the main screen appeared to be a reasonable choice in our usability tests, many of our users in the final user study kept trying to scroll the list up and

down instead of navigating to the "more tags" page. An interesting design space to explore in the future would be ways of showing all tags of the playing song in the main screen while still keeping the cognitive load to a minimum.

Excluding certain types of music:
Some of the users that took part in our usability tests and the final lab study expressed a need for telling the system that they did not want to hear a certain type of music, by excluding a certain tag. The positive nature of tags makes this difficult to support. One solution would be to use methods such as Latent Semantic Analysis to transform the folksonomy into a set of topics [20, 33]. Having this information, one could then define topics as being contrary to each other, and thus utilize the available information on each song to exclude it from a recommended set. This was, however, beyond the scope of TagFlip at this stage.

Integration with conventional applications:
Some of our participants tried clicking on the album art of the current song, expecting to be taken to a dedicated page for the album or artist, and some mentioned they were more interested in artist-based exploration than song-based. An interesting avenue to explore is ways of integrating such functionalities into a tag-based navigation interface, or adding the latter to the more conventional music streaming tools.

Limitations and further future work

Our lab study was intended for understanding the capabilities of TagFlip compared to what people use in their daily lives. A natural next step would be a longitudinal field study, having TagFlip be part of the users' daily music consumption routines for weeks, while studying how it is used and how naturally it fits. The design questions discussed earlier could then be tested for and analysed in a more realistic environment [21].

A potential confounding factor, which is inherent to the methodological choice we took, is the *demand characteristic effect* [13, 24]. This effect describes the possible tendency of participants to subjectively rate our tool higher because they realize we built it. Being aware of this potential threat, we tried to alleviate it by carefully avoiding to bias our participants in any way. For instance, we did not tell the participants that used Spotify first about the second app that they were going to compare it to; they were only told that we were comparing two apps. In addition, the fact that TagFlip did not end up being rated significantly better than Spotify in the more general questions about recommendation quality, can suggest that such an effect was not strong. On the other hand, since most participants had previous experience with Spotify's app, a boost to usability ratings for Spotify is also expected.

The library of music used in TagFlip contained "only" 100,000 songs, and this had tangible negative effects on user experience compared to Spotify with more than 30 million songs. For instance, as our study was performed in Austria, some users pointed to our lack of local songs as an issue. As such, having a larger library and one that better reflects the interests of the target audience can improve user satisfaction.

Another limitation of TagFlip relates to its tag data. Although we meticulously cleaned the Last.fm data from half a million tags to 358, we did not perform any synonym modelling. While this may not be necessary for certain tag types (genre, instrument, etc.) having such a model can improve the experience with mood tags. Moreover, as some songs have few tags, auto-tagging algorithms [10] could be used to propagate more tags to such songs, to get a more uniform library.

CONCLUSION

The primary goal in designing TagFlip was to increase user involvement and control in the process of music recommendation, while keeping the user's mental and interactive effort as small as possible, and fitting the design in a small phone screen. With the positive results of our evaluations, we believe TagFlip has succeeded in its mission. The consensus among our 16 participants of the lab study was that our tool fills an important space that is unsupported by conventional recommender services, and seven of them asked to have the app on their own phones as soon as it was possible. Considering the fact that Spotify's library is more than 300 times larger than ours and its mobile interface has been refined for years, we find it encouraging that TagFlip could perform as well in usability, and come out on top in most recommendation aspects.

In a lab study, we had participants compare TagFlip to Spotify's mobile application, in terms of usability and music recommendation capabilities. Out of our seven hypotheses to test TagFlip, five were supported by our results. In subjective user feedback, these related to (1) aggregate rating in recommendation aspects, (2) interface and interaction adequacy, (3) control and transparency, (4) attitudes and behavioural intentions. The fifth confirmed hypothesis concerned the objective measure of number of interactions per liked song, indicating that TagFlip required less effort from users for discovering a comparable number of new liked songs.

Based on our design process and final evaluation, we reported on a number of design considerations and open design questions regarding tag based music listening and discovery interfaces. Among other things, we found that grouping tags into categories such as genres and moods can help give the user a holistic understanding of the played music; that providing a way to exclude certain types of music based on tags might help enhance user experience; that having a very clear separation between tags of the current song and constraints for the next is crucial; and that facilitating control on strength of tags might improve perceived recommendation quality but comes at the expense of added complexity and confusion.

In future work, we plan to expand our understanding of how TagFlip can fit into the music listeners' daily lives through a long term study. This would also serve as a platform for further studying and comparing alternative solutions to some of the design questions posed earlier. In addition, we intend to improve our library and dataset by including more songs and enhancing our tag space through methods such as auto-tagging and synonym modelling.

REFERENCES

1. Baur, D., Boring, S., and Butz, A. Rush: Repeated Recommendations on Mobile Devices. In *Proceedings of the 2010 International Conference on Intelligent User Interfaces (IUI '10)* (2010), 91–100.
2. Baur, D., Hering, B., Boring, S., and Butz, A. Who Needs Interaction Anyway? Exploring Mobile Playlist Creation from Manual to Automatic. In *Proceedings of the 2011 international Conference on Intelligent User Interfaces (IUI '11)* (2011), 291–294.
3. Bertin-Mahieux, T., Ellis, D. P., Whitman, B., and Lamere, P. The Million Song Dataset. In *Proceedings of the 12th International Society for Music Information Retrieval Conference (ISMIR '11)* (2011), 591–596.
4. Boland, D., Mclachlan, R., and Murray-smith, R. Engaging with Mobile Music Retrieval. In *Proceedings of the 2015 International Conference on Human-Computer Interaction with Mobile Devices and Services (MobileHCI '15)* (2015), 484–493.
5. Bostandjiev, S., O'Donovan, J., and Höllerer, T. TasteWeights: A Visual Interactive Hybrid Recommender System. In *Proceedings of the 2012 ACM conference on Recommender systems (RecSys '12)*, ACM Press (2012), 35–42.
6. Brooke, J. SUS: A Quick and Dirty Usability Scale. Tech. rep., 1996.
7. Burke, R., Hammond, K., and Yound, B. The FindMe Approach to Assisted Browsing. *IEEE Expert 12*, 4 (Jul 1997), 32–40.
8. Chen, L., and Pu, P. Critiquing-based recommenders: Survey and emerging trends. *User Modeling and User-Adapted Interaction 22*, 1-2 (2011), 125–150.
9. Dachselt, R., Frisch, M., and Weiland, M. FacetZoom: A Continuous Multi-scale Widget for Navigating Hierarchical Metadata. In *Proceeding of the 2008 SIGCHI Conference on Human Factors in Computing Systems (CHI '08)* (2008), 1353–1356.
10. Eck, D., Lamere, P., Bertin-Mahieux, T., and Green, S. Automatic Generation of Social Tags for Music Recommendation. In *Advances in Neural Information Processing Systems (NIPS '07)* (2007), 385–392.
11. Greasley, A. E., and Lamont, A. Exploring engagement with music in everyday life using experience sampling methodology. *Musicae Scientiae 15*, 1 (2011), 45–71.
12. Green, S., Lamere, P., Alexander, J., and Maillet, F. Generating Transparent, Steerable Recommendations from Textual Descriptions of Items. In *Proceedings of the 2009 ACM Conference on Recommender Systems (RecSys'09)*, ACM Press (2009), 281–284.
13. Intons-Peterson, M. J. Imagery paradigms: How vulnerable are they to experimenters' expectations? *Journal of Experimental Psychology: Human Perception and Performance 9*, 3 (1983), 394.
14. Jennings, D. *Net, Blogs and Rock 'n' Roll: How Digital Discovery Works and What it Means for Consumers*. Nicholas Brealey Publishing, 2007.
15. Kamalzadeh, M., Baur, D., and Möller, T. A Survey on Music Listening and Management Behaviours. In *Proceedings of the 13th International Society for Music Information Retrieval Conference (ISMIR '12)* (2012), 373–378.
16. Kamalzadeh, M., Baur, D., and Möller, T. Listen or Interact? A Large-Scale Survey on Music Listening and Management Behaviours. *Journal of New Music Research* (in press).
17. Konstan, J. A., and Riedl, J. Recommender systems: From algorithms to user experience. *User Modeling and User-Adapted Interaction 22*, 1 (2012), 101–123.
18. Lamere, P. Social tagging and music information retrieval. *Journal of New Music Research 37*, 2 (2008), 101–114.
19. Lazar, J. *Research Methods In Human-Computer Interaction*. Wiley, 2010.
20. Levy, M., and Sandler, M. Learning Latent Semantic Models for Music from Social Tags. *Journal of New Music Research 37*, 2 (June 2008), 137–150.
21. McGrath, J. E. Methodology matters: Doing research in the behavioral and social sciences. In *Readings in Human-Computer Interaction: Towards the Year 2000*, 2nd ed. Morgan Kaufmann, 1995, 152–169.
22. McNee, S., Lam, S., Konstan, J., and Riedl, J. Interfaces for Eliciting New User Preferences in Recommender Systems. In *User Modeling 2003*, P. Brusilovsky, A. Corbett, and F. de Rosis, Eds., vol. 2702 of *Lecture Notes in Computer Science*. Springer Berlin Heidelberg, 2003, 178–187.
23. O'Donovan, J., Smyth, B., Gretarsson, B., Bostandjiev, S., and Höllerer, T. PeerChooser: Visual Interactive Recommendation. In *Proceedings of the 2008 SIGCHI Conference on Human Factors in Computing Systems (CHI '08)* (2008), 1085–1088.
24. Orne, M. T. Demand Characteristics and the Concept of Quasi-controls. *Artifacts in Behavioral Research: Robert Rosenthal and Ralph L. Rosnows Classic Books* (2009), 110.
25. Oudenne, A. M., Kim, Y. E., and Turnbull, D. S. Meerkat: Exploring Semantic Music Discovery Using Personalized Radio. In *Proceedings of the 11th International Conference on Multimedia Information Retrieval (MIR '10)* (2010), 429–432.
26. Pampalk, E., and Goto, M. MusicSun: A New Approach to Artist Recommendation. In *Proceedings of the 8th International Conference on Music Information Retrieval (ISMIR '07)* (2007).

27. Pampalk, E., Rauber, A., and Merkl, D. Content-based Organization and Visualization of Music Archives. In *Proceedings of the 10th ACM International Conference on Multimedia (MULTIMEDIA'02)* (2002), 570–579.

28. Pu, P., and Chen, L. Trust building with explanation interfaces. In *Proceedings of the 15th International Conference on Intelligent User Interfaces (IUI'06)* (2006), 93–100.

29. Pu, P., and Chen, L. A user-centric evaluation framework of recommender systems. In *UCERSTI Workshop of RecSys'10* (2010), 14–21.

30. Schwartz, B. *The Paradox of Choice*. Harper Perennial, 2004.

31. Schwartz, B., Ward, A., Monterosso, J., Lyubomirsky, S., White, K., and Lehman, D. R. Maximizing Versus Satisficing: Happiness Is a Matter of Choice. *Journal of Personality and Social Psychology 83*, 5 (2002), 1178–1197.

32. Sinha, R., and Swearingen, K. The Role of Transparency in Recommender Systems. In *Proceedings of the 2002 SIGCHI Conference on Human Factors in Computing Systems (CHI' 02)* (2002), 830–831.

33. Sordo, M., Gouyon, F., Sarmento, L., Celma, O., and Serra, X. Inferring Semantic Facets of a Music Folksonomy with Wikipedia. *Journal of New Music Research 42*, 4 (2013), 346–363.

34. Spotify Android SDK. `https://developer.spotify.com/technologies/spotify-android-sdk/`.

35. Stumpf, S., and Muscroft, S. When Users Generate Music Playlists: When Words Leave off, Music Begins? In *Proceedings of the 2011 International Conference on Multimedia and Expo (ICME '11)* (2011), 1–6.

36. Sturm, B. L. The State of the Art Ten Years After a State of the Art: Future Research in Music Information Retrieval. *Journal of New Music Research 43*, 2 (2014), 147–172.

37. Tintarev, N., and Masthoff, J. A Survey of Explanations in Recommender Systems. In *Proceedings of the 2007 IEEE International Conference on Data Engineering Workshop* (2007), 801–810.

38. Tzanetakis, G., and Cook, P. MARSYAS3D: A Prototype Audio Browser-editor Using a Large Scale Immersive Visual and Audio Display. In *Proceedings of the 2011 International Conference on Auditory Display* (2001), 250–254.

39. van Gulik, R., Vignoli, F., and van de Wetering, H. Mapping Music in the Palm of Your Hand, Explore and Discover Your Collection. In *Proceedings of the 5th International Society for Music Information Retrieval Conference (ISMIR'04)* (2004).

40. Verbert, K., Parra, D., Brusilovsky, P., and Duval, E. Visualizing Recommendations to Support Exploration, Transparency and Controllability. In *Proceedings of the 2013 International Conference on Intelligent User Interfaces (IUI '13)*, ACM Press (2013), 351–361.

41. Vig, J., Sen, S., and Riedl, J. Tagsplanations: Explaining Recommendations Using Tags. In *Proceedings of the 2009 International Conference on Intelligent User Interfaces (IUI '09)* (2009), 47–56.

42. Vig, J., Sen, S., and Riedl, J. Navigating the Tag Genome. In *Proceedings of the 2011 International Conference on Intelligent User Interfaces (IUI '11)*, ACM Press (2011), 93–102.

43. Vignoli, F. Digital Music Interaction Concepts: A User Study. In *Proceedings of the 5th International Society for Music Information Retrieval Conference (ISMIR '04)* (2004), 415–420.

44. Wang, J.-C., Shih, Y.-C., Wu, M.-S., Wang, H.-M., and Jeng, S.-K. Colorizing Tags in Tag Cloud: A Novel Query-by-Tag Music Search System. In *Proceedings of the 2011 ACM International Conference on Multimedia (MM '11)*, ACM Press (2011), 293–302.

45. Wang, J.-C., Wu, M.-s., Wang, H.-m., and Jeng, S.-k. Query by Multi-tags with Multi-level Preferences for Content-based Music Retrieval. In *Proceedings of the 2011 IEEE International Conference on Multimedia and Expo (ICME '11)* (2011), 1–6.

46. Yee, K.-P., Swearingen, K., Li, K., and Hearst, M. Faceted metadata for image search and browsing. In *Proceedings of the 2003 SIGCHI Conference on Human factors in Computing Systems (CHI '03)* (2003), 401–408.

47. Zhu, S., Cai, J., Zhang, J., Li, Z., Wang, J.-c., and Wang, Y. Bridging the User Intention Gap: An Intelligent and Interactive Multidimensional Music Search Engine Categories and Subject Descriptors. In *Proceedings of the 2014 ACM International Workshop on Internet-Scale Multimedia Management (WISMM '14)* (2014), 59–64.