Semester Project Report

Validation of Game Theoretic Algorithms for Group Recommender Systems

Coordinator teacher:
Prof. Dr. Pearl Pu

PhD Student:
Popescu George

EPFL, September 18, 2016

Abstract

In online group and social recommender systems algorithms play the essential role of aggregating users' preferences. Such systems act as decision support frameworks for users to negotiate on a given outcome they like to consume with other members. Users are seldom satisfied by one recommended outcome. They prefer algorithms which are transparent and generate dynamic recommendations adapted to their changing preferences. Furthermore, they expect that a voting aggregation mechanism will satisfy everyone's tastes and propose best items for the group.

In this report we consider the case of online group decision in the music domain. We present our pilot-test results with respect to the validation of 4 algorithms implemented in the GroupFun recommender system: probabilistic weighted sum, deterministic weighted sum, least misery and random selection. The main results point to the fact that individuals enjoy discovering their friends' music tastes and give ratings according to their own judgments. They are sensitive to the recommendations made for the entire group and interpret them at a personal level. Users prefer a selection mechanism that favors "democracy" and everyone's involvement. Moreover, our interviews outline user expectations in terms of: high enjoyability while interacting with other users via the system, excellent aggregation transparency and easy to understand recommendations based on the group's ratings.

1. INTRODUCTION

In group recommender systems algorithms are used to help users identify most relevant items. They implement various aggregation strategies which are used to suggest top k best items from a lengthier set. These strategies aim at increasing the group's welfare. Most common used deterministic strategies are presented in (Masthoff, 2005): plurality voting, utilitarian, approval voting, least misery, most pleasure, average without misery, fairness, most respected person, Borda count, Copeland rule or Kemeny scores. Due to the increased complexity of some of them, only simple to understand algorithms are implemented in popular group recommender systems: e.g. PolyLens uses the least misery strategy. Furthermore, individuals use simple strategies mentioned above when judging different outcomes, particularly the average strategy, the average without misery and the least misery strategy. However, there is no clearly dominant strategy. Fairness plays an important role in group preference aggregation, but subjects do not have a clear strategy for applying it and expect that the system should take this into account (Masthoff, 2004).

In this report we study "which group algorithm best satisfy users expectations?" by comparing 4 algorithms: probabilistic weighted sum, deterministic weighted sum, least misery and random. The present report also investigates upon:

1. Which algorithm is best suited to meet users' expectations?
2. How users perceive the algorithms accuracy and fairness?

3. How users perceive the 4 recommendations given to them?
4. How users like to negotiate in a group?

2. ALGORITHMS

We consider the music domain in which many users usually form many groups and listen to many songs. Given the fact that the length of one song is of 3 to 4 minutes users usually select a playlist containing several to lots of songs. Thus, this domain presents both opportunities and challenges since the recommendation needs to focus on both diversity and accuracy.
We propose the following 4 algorithms for comparison:

- PS (Probabilistic Sum): select the common playlist' songs probabilistically, each of them having the same probability to be selected
- LM (Least Misery): select songs with the highest minimum individual ratings
- DWS (Deterministic Weighted Sum): deterministically select songs with the highest score
- PWS (Probabilistic Weighted Sum): compute weighted sum and select songs based on their score probabilities.

2.1. GENERAL FRAMEWORK

Let A be the set of all users and S the set of all possible outcomes that can be rated. In our group music recommendation setting, the outcomes are songs s_i that are selected in the common playlist. We let each user a_j submit a numerical vote $score(s_i, a_j)$ for each song s_i that reflects their preference for that song. These votes are given as ratings on a 5-point Likert scale and normalized so that the scores given by each user sum to 1:

$$score(s_i, a_j) = \frac{rating(s_i, a_j)}{\sum_i rating(s_i, a_j)} \quad (1)$$

We then assign a joint score to each song that is computed as the sum of the scores given by the individual users:

$$score(s_i) = \sum_{a_j \in A} score(s_i, a_j) \quad (2)$$

To choose the songs to be included in a playlist of length k, a deterministic method is to choose the k songs with the highest joint rating: weighted sum (DWS):

$$score(s_i) := {score(s_i)} \Big/ {\sum_{si \in S} score(s_i)} \quad (3)$$

The probabilistic weighted sum (PWS) iteratively selects each of the k songs randomly

according to the probability distribution:

$$p(s_i) = {score(s_i)} \Big/ {\sum_{si \in S} score(s_i)} \quad (4)$$

By comparison, the probabilistic sum (PS) method chooses the k songs with equal probability:

$$p(s_i) = {1} \Big/ {|S|} \quad (5)$$

The least misery (LM) method takes into account the minimum of ratings for each user:

$$\min(score(s_i, a_j)), \forall a_j \in A \quad (6)$$

2.2. EXAMPLE

To illustrate how each algorithm works, we consider the following example. In the next table, user1, user2, and user 3 represent group members. The score distribution normalized to 1 for each of the users is displayed in the respective row, and the joint scores are shown in the table below.

Table I. Item selection example using the 4 algorithms

User1	Song1: 0.1	Song2: 0.4	Song3: 0.4	Song4: 0.1
User2	Song1: __	Song2: 0.2	Song3: __	Song4: 0.8
User3	Song1: 0.4	Song2: 0.2	Song3: __	Song4: 0.4
Total score	Song1: 0.5	Song2: 0.8	Song3: 0.4	Song4: 1.3

The least misery (LM) will choose song 2 and song 3 (each of them has the minimal rating 0.2). For all other songs the minimum score is 0.1. After normalizing the total scores by the sum of scores, we obtain the following probability distribution for the set of outcomes.

Table II. Probability distribution

P	Song1: 0.16	Song2: 0.26	Song3: 0.13	Song 4: 0.43

Considering the probability as the final score, the deterministic weighted sum (DWS) will chose songs 4, 2, 1 and 3. Probabilistic weighted sum (PWS) will choose one song after another using this probabilistic distribution. Compared to other social choice based algorithms, PWS is incentive compatible. That is, it is to the best interest of the individual to reveal his/her preferences truthfully. It is in fact equivalent to a random dictator method, where the dictator will choose a song randomly with the probabilities given by its degree of preference – a reasonable method since nobody wants to hear the same song

over and over again. This is because the probability of a song s_i to be chosen can be written as:

$$p(s_i) = {score(s_i)}\Big/{|A|} = \sum_{a_j \in A} \frac{1}{|A|} score(s_i, a_j) \quad (7)$$

or, in other words, the probability of choosing user a_j times the normalized score that user a_j has given to song s_i. Indeed, User3's preference for song 1 yields a significant probability that this song will be included in the playlist, relative to other songs.

2.3. DISCUSSION

The contribution of the PWS algorithm in the paper stands out with respect to group satisfaction. We expect users to be more satisfied using PWS than other algorithms given their expectations to discover the music of other members.

Advantages of PWS compared with the other algorithms:

1. Users are free to choose the number of songs
2. Ratings are updated permanently
3. The algorithm is computationally simple
4. Users can negotiate their ratings and trade utility
5. Incentive-compatible truthful property is observed
6. The algorithm favors music diversity

The disadvantages of PWS are:

1. It is difficult to quantify rating differences between distinct users. The weights given by each user cannot be compared with the ones given by another since users have different estimations of their utility.
2. Self-selection effect: most popular songs will receive most votes - not ideal if long tail distribution is desired.

Since PWS can be interpreted as similar to the random scheme users have to test it in more recommendation rounds to understand its inner logic. PWS can be further developed to include the group dynamics. One solution is to consider trust and other members' comments on the songs rated by one user (e.g. "like"/"dislike").

3. GROUPFUN INTERFACE

GroupFun is a web application that helps a group of friends to agree on a common music playlist for a given event they will attend, e.g. a graduation ceremony. Firstly, it is implemented as a Facebook plugin connecting users to their friends. Secondly, it is a music application that helps individuals to manage and share their favorite music with groups. In GroupFun users can listen to their own collection of songs as well as their friends' music. With the collective music database, the application integrates friends' music tastes and recommends a common playlists to them. Therefore, the application aims at satisfying music tastes of the whole group by aggregating individual preferences through the use of previously presented algorithms.

Figure 1. "Home" page of GroupFun

In the "Home" page users see 4 playlists: one from a recent event, one containing popular songs, one from a party and the last one from an older event. They can listen to each song in each of the playlists.

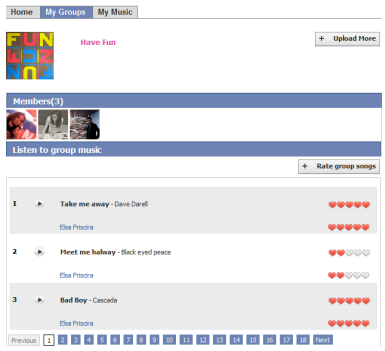


Figure 2. "My Groups" page of GroupFun

In the "My Group" page users can create groups, upload and rate their music, invite friends and hear the group's songs. Finally, in the "My Music" page users see their contribution to GroupFun: for each song is displayed the associated group, the rating and its name and artist. Users can also listen to their individual preferences in the same interface. One of the most important characteristics of GroupFun is that it combines

music, friends and groups together. In other words, GroupFun serves as a 3-in-1 magic box that allows users to conveniently organize their individual music library, effectively communicate with friends and actively participate in social activities.

4. EXPERIMENT DESIGN

In this chapter we highlight our experiment analysis results based on a pilot-study design. We start from initial criteria, user roles and present the overall experiment description. Then we present our questionnaire results. Lastly we present and discuss the main suggestions pointed out by our interviewees during the experiment.

There are 4 algorithms and 1 interface. Each of the algorithms displays a common list of 15 songs to all group members.

The final outcome is displayed to all group members after they have finished their tasks.

Table III. Evaluation of the 4 algorithms with the same interface

Interface/Algorithm	**PWS=Alg1**	**DWS=Alg2**	**LM=Alg3**	**Rand=Alg4**
GroupFun	8	8	8	8

We follow a with-in group experiment design. For each cell in the table above we recruited 8 users. All users tested the results of all 4 algorithms.

4.1. CRITERIA

We identified the following criteria for selecting users to join GroupFun: a group of friends connected on Facebook prepared an individual collection of songs in .mp3 format on their computers.

4.2. INCENTIVES

The most important incentive to use GroupFun was that users could save their personal music in the system and listen to group music anytime from their personal devices.

4.3. SCENARIO

We developed the following scenario that users had in mind when using GroupFun: 2 groups of 4 users are driving from Lausanne to Geneva and listen to group music together (around 45 min's drive). The assumption for selecting this scenario is that users only need a small number of songs so that we recommend only 15:

15 songs x 3 minutes / song = 45 min's total listening time

4.4. ROLES

Each of the 4 members of each group had 1 of the following roles, each having one distinctive identity label:

- 1 driver
- 3 passengers

4.5. MOTIVATION

Users upload their own favorite music given the fact that they will like to share their music with their friends for the drive instead of listening to random music from the radio stations.

4.6. MATERIALS

The below-listed materials were used during the experiment:

- Hardware:
  - users' personal computers
  - laboratory computers
- Software:
  - GroupFun application
  - Online evaluation questionnaire
  - Evaluation notes
  - Free YouTube to Mp3 converter for mp3 files download

4.7. HYPOTHESES

We present the following hypotheses prior to our experiment based on the algorithms' properties.
H1: Users do not take into account other members' preferences when initially eliciting individual preferences.
H2: Users like to know other members' music preferences.
H3: Users consider other group members' preferences when they are reviewing recommended songs.
H4: Users prefer PWS to DWS given the increased diversity of the recommendations.

4.8. USERS' TASKS

Users evaluated the recommended songs. They were free to interact with the system following a series of steps displayed to them. Each subject evaluated all 4 algorithms. First, all users were asked to upload their music and to give ratings to their songs. They used Free YouTube to Mp3 Converter (http://www.dvdvideosoft.com/products/dvd/Free-YouTube-to-MP3-Converter.htm) to download some .mp3 music files they like from YouTube. This process took some time since the download time cannot be shorter than the listening time. Consequently, some users expressed their wish to upload more songs but due to time constraints we had to limit this. After selecting all of his/her individual songs, the driver proceeded to create groups and invite his/her friends to join. 4 interfaces

in GroupFun were used to display the final outcomes of the 4 algorithms. Finally subjects were asked to fill in their subjective responses in an online evaluation questionnaire considering the output of all algorithms.

1) Meeting time and place: we asked each group to come to a meeting area in BC 144.

2) Instructions to the users: the experiment's administrator (hereafter admin) first debriefs each group on the nature of the experiment. He told users about the purpose of the experiment: study of algorithms and user issues. Users did not discuss with each-other until the outcome negotiation at the end. The admin explains to all 4 users the overall functionalities provided in GroupFun, e.g.: listen to the music from the "Home" page. For each group the "driver" creates a group, invites friends and uploads music. Other friends accept the invitation, upload and rate their songs. Users were told to pay attention to the songs recommendation for the whole group. Optionally they could upload more songs and/or change personal ratings at any time. When they finished uploading and rating they were instructed to evaluate the songs recommended by the 4 algorithms separately. After this step 5 minutes was offered to all group members to negotiate for arriving at a final outcome as common list of songs. Finally an online questionnaire and interview session were scheduled to assess users' perceived satisfaction.

3) Start using GroupFun: the admin assists users with all technical details for the steps mentioned above.

The driver tasks:

1. Login to GroupFun using Facebook (apps.facebook.com/groupfun). Alternatively use grpupc1.epfl.ch/~laurentiu/groupfunElsaPhpFram/index.php
2. Create a group
3. Invite friends
4. Upload music
5. Rate music
6. Upload and rate more songs (optional)

The passengers' tasks:

1. Receive the invitation
2. Join the group
3. Upload music
4. Rate music
5. Upload and rate more songs (optional)

4) Group negotiation on a list of songs.

5) Questionnaire (http://tiny.cc/y6w0x)

To conclude the study, survey questions were used in order to assess the users' experience of using the system. The questionnaire items were statements for which a user indicated his/her level of agreement on a five-point Likert scale ranging from 1 to 5, where 1 means "strongly disagree" and 5 means "strongly agree".

I am satisfied with the songs recommended to me by: *

	strongly disagree = 1	2	3	4	5 = strongly agree
Alg1	○	○	○	○	○
Alg2	○	○	○	○	○
Alg3	○	○	○	○	○
Alg4	○	○	○	○	○

Figure **3**. Algorithms evaluation question

6) Interviews

During the 15 minutes interview we had an open discussion with the users. This part was very useful since users expressed freely how they perceived both advantages and drawbacks of the 4 algorithms. With the results of the experiment we were able to judge the influence of each recommendation algorithm on users' perceived satisfaction.

5. RESULTS

8 participants (1 female) were recruited to participate in the experiment. They were Master (5) and PhD (3) students at EPFL. Their ages range from 20 to 29 (1 person 20-24) and had various nationalities (Romanian, Greek, Italian, French, etc.). All subjects had solid IT experience and frequently used music players (such as iTunes, Winamp, Windows Media Player, last.fm, Deezer.fr, etc.). In a typical week 6 individuals used music applications less that 5h per week, 1 between 5 and 10 hours and 1 between 10 to 15h per week.

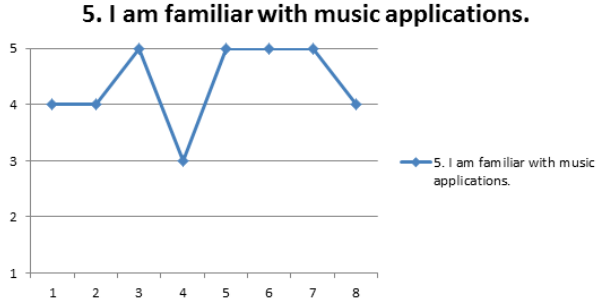


Figure 4. User familiarity of music applications

An example of one user ratings is given below. The scale is from 1 to 5 and ratings are given to group songs from 1 to 7.

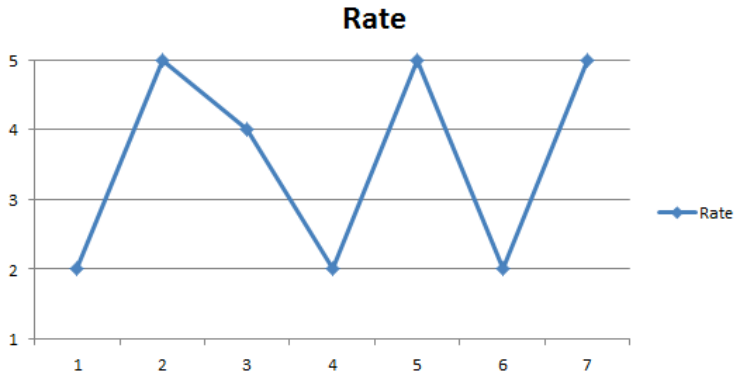


Figure 5. Example of user ratings

For a different user we notice a different trend in giving song ratings. The two charts show that users analyzed each song separately and gave ratings according to their own valuations.

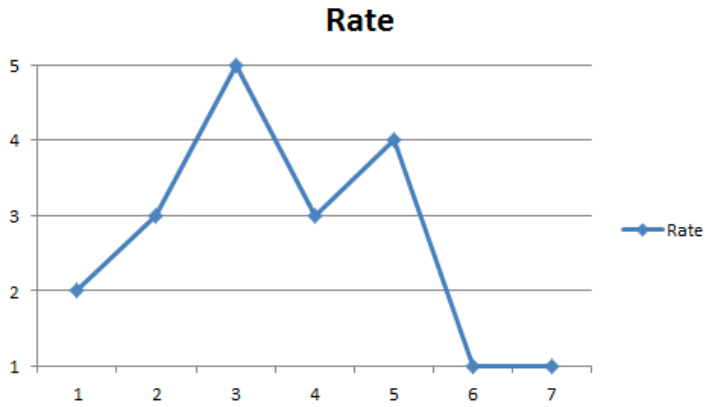


Figure 6. Another example of user ratings

5.1. GROUP1 DETAILS

The first driver created a group named: "**Road Trip 2**" and invited 3 friends. In total a number of **19 songs** were uploaded. For a total of **8 songs** all users gave ratings. Other songs received ratings for fewer friends.

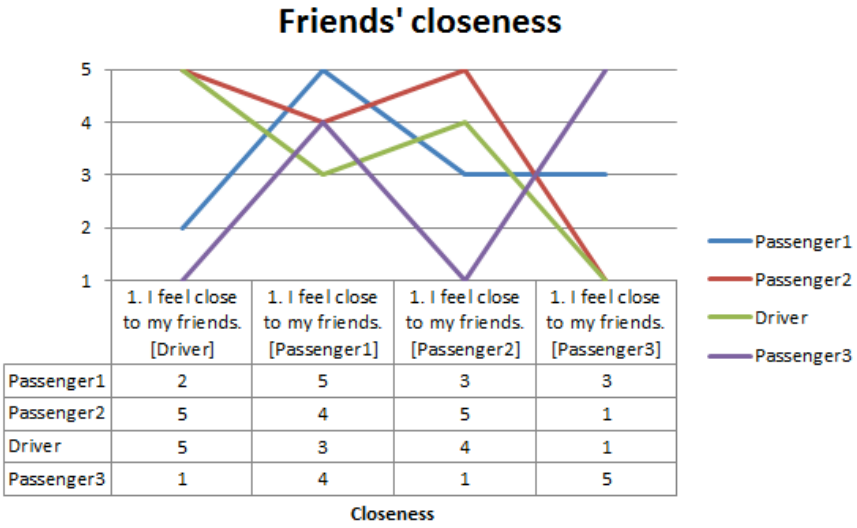
Table **IV**. Number of songs uploaded per user – 1st group

	User1	User2	User3	User4
# songs	2	6	4	7

A print-screen of user ratings in Group1 is available below.

Figure 7. Users' ratings in Group1

We asked the first group "how close they feel to their friends". In the chart below each of the 4 users gave a rating from 1 to 5 as a measure of their closeness to their friends. We observe that Passenger 3 feels only close to Passenger1 whereas the Driver feels close to Passenger1 and Passenger2 and not close at all to Passenger3.



	1. I feel close to my friends. [Driver]	1. I feel close to my friends. [Passenger1]	1. I feel close to my friends. [Passenger2]	1. I feel close to my friends. [Passenger3]
Passenger1	2	5	3	3
Passenger2	5	4	5	1
Driver	5	3	4	1
Passenger3	1	4	1	5

Figure 8. Friends' closeness one-another in Group1

When asking the user if they know well their friends' musical tastes we received very different answers in group1. For instance Passenger3 reported that he does not know at all his friends' musical tastes. On the opposite side the driver knows well Passenger1's and Passenger2's musical tastes. The chart also shows that only Passenger1 knows Passenger3's musical tastes to some extent.

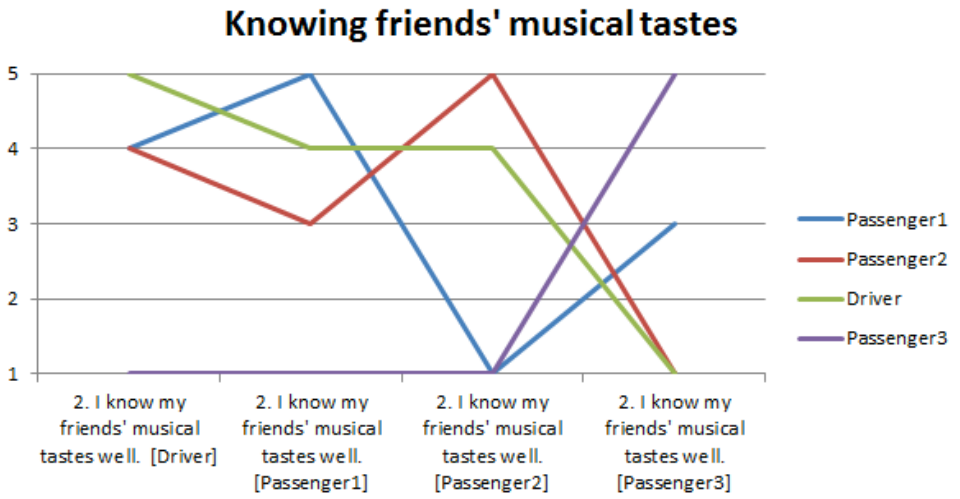


Figure 9. Friends' musical tastes in Group1

5.2. GROUP2 DETAILS

The second driver created a group named: "**CrazyGroup**" and invited 3 friends. In total a number of **19 songs** (again) were uploaded. In this group all 4 users rated all of the group songs.

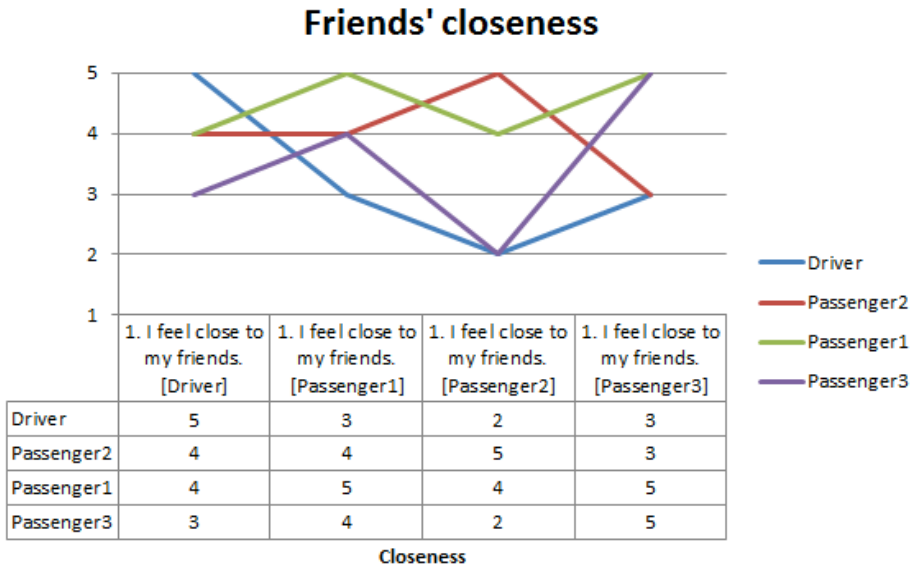
Table **V**. Number of songs uploaded per user – 2st group

	User1	User2	User3	User4
# songs	3	9	6	1

A print-screen of user ratings in Group2 is available below.

Figure 10. Users' ratings in Group2

We asked the second group the same question about closeness to friends. We report that in this group individuals feel closer one-another. For instance Passenger1 mentioned he feels very close to all other group members (green line). Passenger3 only feels close to the Driver and Passenger1 and not that close (score of 2) to Passenger2. The driver gave the highest score of 3 to Passenger1 and Passenger3 and only a score of 2 to Passenger2.



	1. I feel close to my friends. [Driver]	1. I feel close to my friends. [Passenger1]	1. I feel close to my friends. [Passenger2]	1. I feel close to my friends. [Passenger3]
Driver	5	3	2	3
Passenger2	4	4	5	3
Passenger1	4	5	4	5
Passenger3	3	4	2	5

Figure 11. Friends' closeness one-another in Group2

In group2 there is a high heterogeneity of knowledge of users' musical tastes. The driver knows relatively well the others' musical tastes whereas Passenger2 reported only a score of 2 to for other group members. Passenger3 also knows relatively well The Driver's and Passenger1's musical tastes.

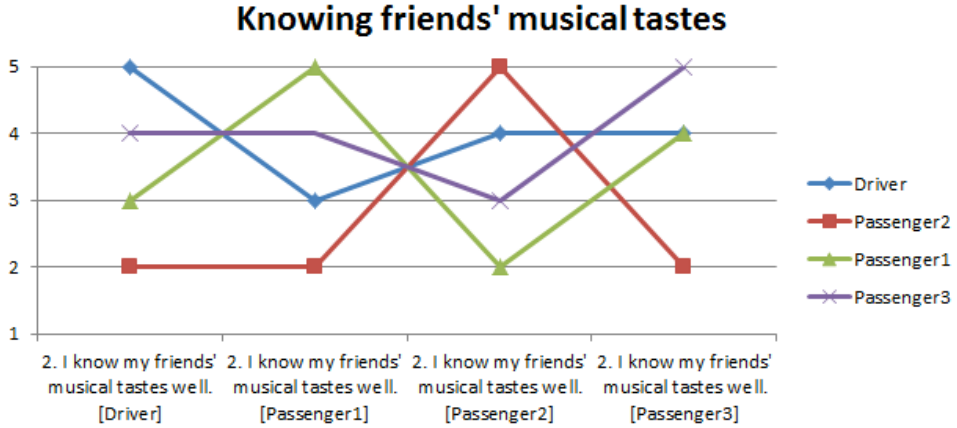


Figure 12. Friends' musical tastes in Group2

5.3. QUESTIONNAIRE RESULTS

5.3.1. FACTORS FOR A GOOD PLAYLIST

In replying to "which factors are most important for a good playlist?":

- **2** users chose "Many of the songs are suggested by me"
- **4** users → "The songs in the playlist are diverse"
- **All 8** users → "The playlist helps me discover new music"
- **4** users → "The playlist has a good transition between songs"

1. I consider the most important factor(s) for a good playlist to be (check as many as you want):
The songs in the playlist are diverse, The playlist helps me discover new music
The playlist helps me discover new music
The songs in the playlist are diverse, The playlist helps me discover new music, The playlist has a good transition between songs
The songs in the playlist are diverse, The playlist helps me discover new music, The playlist has a good transition between songs
The playlist helps me discover new music, The playlist has a good transition between songs
The songs in the playlist are diverse, The playlist has a good transition between songs
Many of the songs are suggested by me, The playlist helps me discover new music
Many of the songs are suggested by me, The playlist helps me discover new music

Figure 13. Important factors for a good playlist

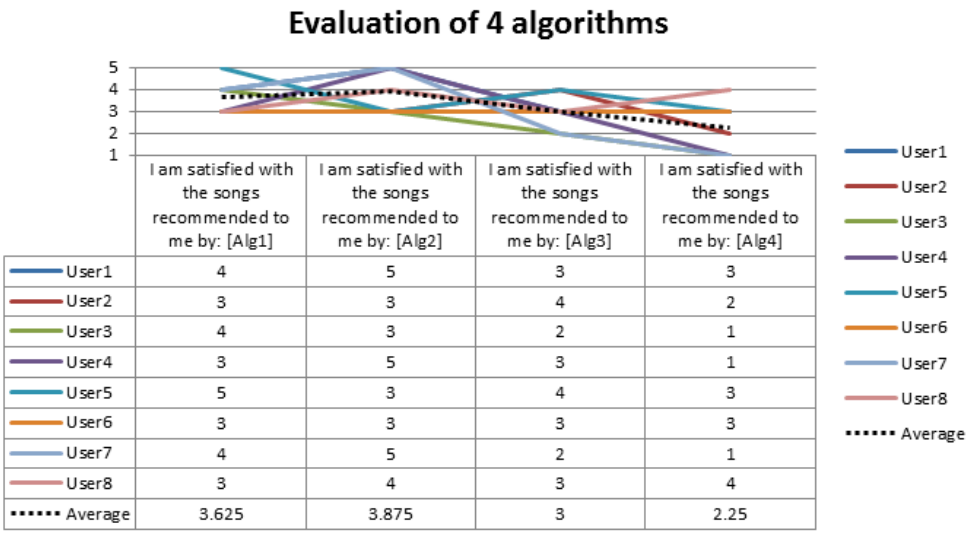
5.3.2. EVALUATION OF THE 4 ALGORITHMS

We implemented:

- **PWS as Alg1**
- **DWS as Alg2**
- **LM as Alg3**
- **PS as Alg4**

The results presented in the chart and table below are very exciting for our development of the probabilistic weighted sum algorithm. With colored lines are presented all of the 8 users' ratings and with a dashed black line the average of all results. We notice a favorable trend for the first two algorithms. The least misery one is less preferred in general by all members compared with the first two whereas the random or probabilistic selection received the least scores.

The last row in the table shows that the average scores for PWS and DWS are very close: 3.625 for the first one and 3.875 for the second one. Given the fact that in our experiment users did not have the time to experience the advantages of PWS in many voting sessions we find this result very encouraging for future research. Moreover, we note that none of the users gave a score lower than 3 (out of 5) for PWS and DWS. Not so good results are highlighted for the last two algorithms. Users reported that they noticed randomness in the results of last algorithm meaning that it actually does not take into account their ratings.



	I am satisfied with the songs recommended to me by: [Alg1]	I am satisfied with the songs recommended to me by: [Alg2]	I am satisfied with the songs recommended to me by: [Alg3]	I am satisfied with the songs recommended to me by: [Alg4]
User1	4	5	3	3
User2	3	3	4	2
User3	4	3	2	1
User4	3	5	3	1
User5	5	3	4	3
User6	3	3	3	3
User7	4	5	2	1
User8	3	4	3	4
Average	3.625	3.875	3	2.25

Figure 14. Evaluation of algorithms

5.3.3. NUMBER OF SONGS

When asked if "the system suggested an adequate number of songs" most users strongly agreed that for the current scenario the number of songs was well chosen. Some of them mentioned that some songs may be longer than 3 minutes (maybe 4 or 5) and that a smaller number (like 10 or 12 songs) may fit better our scenario.

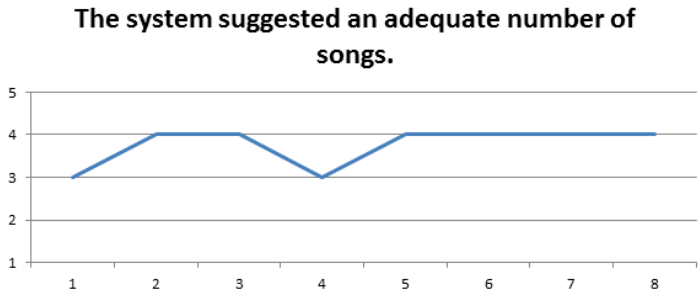


Figure 15. Adequate number of songs

5.3.4. PLAYLIST AGREEMENT

We asked our users if "The playlist is suitable for the whole group." Only one user did not agree to some extent with the playlist because of distinct musical tastes. Other 3 users favored the playlist while 4 others reported that "It's ok!".

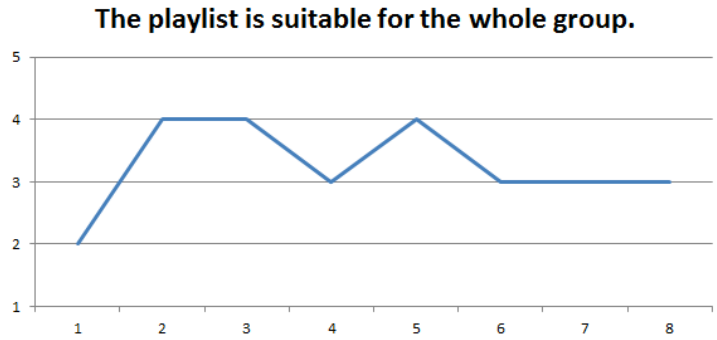


Figure 16. Playlist agreement

5.3.5. FACTORS FOR ACCEPTING A NEW SONG

In replying to which factors are most important for accepting a song which is not uploaded by me a current user:

- **Almost all users (7)** chose "The song match my interest"
- **Only 2** users → "One of my friends likes the song"
- **Only 2** users → "Most of my friends like the song"
- **3** users → "The song match the context"

5. I consider the most important factor(s) for accepting a song which is not uploaded by me to be:
The song match my interest, One of my friends likes the song
The song match my interest, The song match the context
The song match my interest, The song match the context
The song match my interest
The song match my interest, The song match the context
One of my friends likes the song, Most of my friends like the song
The song match my interest
The song match my interest, Most of my friends like the song

Figure 17. Important factors for accepting a new song

5.3.6. INFLUENCE OF OTHER MEMBERS

Another question we formulated to our users was to which extent "my friends' ratings influence mine". As we noticed from the way users gave ratings and how they interpreted the 4 algorithms users judge each individual song according to their own perceptions. Two users are greater influence by their friends' ratings than others. Only 1 user was not at all influenced by others ratings and gave all songs only the score that he appreciated. This result encourages us to carry on future experiments on user interaction and decision making in order to analyze in more detail how users influence each other through both ratings and face-to-face discussions.

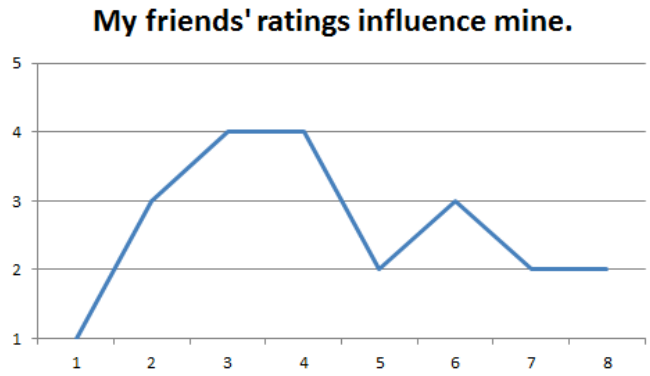


Figure 18. Influence of other members' ratings

5.3.7. USEFULNESS OF GROUPFUN

Towards the end of our online questionnaire we asked our users how useful they find GroupFun. The results we report are extremely pleasant. All of the users agreed that our system is very useful for easing their decision making process. They mentioned to us that it saves them a lot of time to agree on a common playlist they would consume together in a group. In the same time GroupFun proposes a democratic preference elicitation method in which everyone is involved.

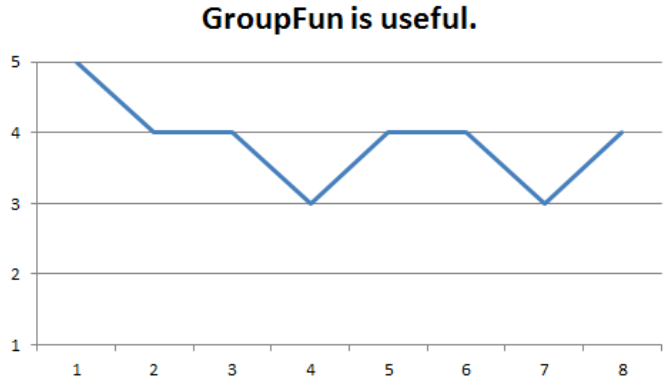


Figure 19. Usefulness of GroupFun

5.3.8. FUTURE USE OF GROUPFUN

Even though our system allows users to express their preferences clearly and interact with other group members some users reported that they are tempted to use GroupFun in the future only given future improvements of both the algorithms and the interface. Most of them replied that the system has a nice implementation at the moment but they also expect future progress.

Figure 20. Future use of GroupFun

5.3.9. TELL FRIENDS ABOUT GROUPFUN

6 of 8 users mentioned that they will tell their friends about GroupFun. They seemed very excited to create new groups with other friends also.

Figure 21. Some users were very excited of telling their friends about groupFun

5.3.10. LISTEN TO GROUP MUSIC ALONE

Part of our users enjoyed listening to the music in GroupFun mostly with their group. They mentioned that they would listen to their own collection of music when they are alone. In the same time they look forward to creating other groups and discovering their friends music tastes through GroupFun. Users were also thinking of uploading other music next time with other friends. Altogether, the GroupFun music is mostly enjoyed within a group rather than alone.

Figure 22. Users are inclined to listen to group music more in groups than alone

5.4. INTERVIEW RESULTS

In the first part of the interview we aimed at finding out specific comments regarding the 4 outcomes.

5.4.1. MANIPULATION

Being asked if they "would you like to change the final outcome provided by different algorithms" some users replied yes, others said that they are not that much interested in changing the outcome but are comfortable with the fact that an algorithm will simply aggregate all preferences together. The ones who did not like the final outcome suggested aligning or competing with others for final selection. Some individuals mentioned that manipulation depends on the mood. Some were very attentive to each song and listened to all of their friends' uploaded songs while others had only an overview of the common list. For the first case they suggested that the evaluation should be done by every user for every song after listening to all group songs. However, they agreed that this process is very time consuming.

In Group2 we found out that when seeing some recommendation results for the first time users generally liked the music because it is new and comes from their friends. They do not "really" want to manipulate. They prefer to state what they like and judge the songs according to their own tastes. The majority of users like to see good group songs on top of the playlist even though there are not their uploaded songs. They agree broadly with the order proposed by the algorithms. Many users tried to chance the ratings of their songs and observe how the playlist changes thus trying to understand the system and manipulate. Some others decided to upload more songs and give higher ratings to personal songs. Since they did not fully understand how the algorithm works they could not observe a significant improvement of their decision.

5.4.2. FAIRNESS

We wanted to know of individuals consider the outcomes of the 4 algorithms to be fair for them. They generally considered to have received a fair recommendation except for random generation. They analyzed the ratings and identified that some algorithms do not take into account, deterministically, all the ratings of all users. 1 person identified that 2 algorithms out of 4 produced random outcomes. He mentioned that the 1st algorithm (PWS) has the most sense by evaluating "some sort of average" of all recommendation lists. In addition he mentioned that top songs were the ones with highest ratings.

In the second group users reported that they did not have enough time to analyze fairness. They generally trusted the system and were confident it produced good results. They were convinced that everyone was threat equally. In the same time they mentioned a "strange change" in some algorithms even though ratings were not different. In this case they could not say if the change came from the algorithm itself or from the fact that some group members changed their ratings or uploaded more songs. By consequence they had to fully trust the algorithm and consider that it took fairness into account to some extent.

5.4.3. INNER LOGIC

One question of our interview was related to users understanding of the inner logic of the system: how it recommends and computes top songs for the group.

Users perceived some outputs to be random and some others to be deterministic. However, they admitted that they did not know in detail how the preference aggregation rule worked. Most users did not have enough time nor were interested in understanding in great detail the inner logic of the system. They simply enjoyed giving ratings and stating their preferences after listening to their friends' songs. Since they only had time to upload 19 songs and the algorithms considered 15 they were only influenced by the order of the songs and aggregated ratings rather than the elimination of some of their songs (4). 1 user mentioned that the 1st algorithm (PWS) computed "some weighted-sum" of ratings and that another one takes "lest ratings into account" (misery).

The second group did not understand how recommendations are generated almost not at all. Individuals believed that some sort of average is computed but sometimes that average rule is not obeyed. Users mentioned that some randomization may be useful to display songs from different users so that they will receive serendipitous recommendations. However this makes the inner logic of the system hard to understand. Some users said that a simple deterministic method should be enough to represent the group preference. Between deterministic and non-deterministic they would choose the later motivated by always knowing the rule which governs how the system works thus enhancing its transparency. Some users paid more attention to simple aggregation rules such as average and identified that one algorithm relatively computes this. In general they

did not analyze the randomness of some algorithms or how much the recommendation varies with time.

5.4.4. DIVERSITY AND PERSONAL INTEREST

We wanted to know if people perceived that they received diverse recommendations, if they were familiar with them, if they matched their interest and how enjoyable they found the recommended music.

The first group perceived the system to recommend songs in a diverse order - e.g. without taking into account music genres. The 1st algorithm produced most diverse recommendations. The whole group agreed that they were familiar to only 30% of the group songs and 70% were unknown to them. Furthermore, they enjoyed a lot all the group songs and added that they matched their interest to a high extent: 3 of 4 users enjoyed discovering their friends' music. 1 outlier user mentioned that he enjoyed only a small part of uploaded songs. Overall, they perceived a great diversity of songs and tastes.

For the second trial users agreed that they received quite diverse recommendations also. Some mentioned than they did not know 30% of the group uploaded songs while for other the percentage raised to 60%. They added that "some music was quite good" and that most songs matched their interest. Only a small proportion was not agreeable at all (10%-20%). One group said that the first algorithm produces the most diverse recommendations and they strongly preferred it. For some people no more than 3 songs in the playlist were new whereas for others at least 25% they never heard before. Since time was short users could only listen to some parts of the songs. They shuffled the songs content and then they gave each song a rating.

5.4.5. NEGOTIATION

Lastly, we were interested in finding how people prefer to discuss or negotiate about the outcomes in a group?

Democratic voting was the choice of the first group. Interviewees agreed to keep the voting mechanism as simple as possible, maybe even reduce it to "like" and "not-like" meaning +1 votes for each songs. In this scenario all users are equal. 1 user suggested setting a limit of songs per user (e.g. max 5 songs to upload). They were fond of discovering their friends' music tastes even though uploaded songs were not very popular. The system offered them a very useful decision support framework. Users preferred to simply list to all 15 songs without thinking of the order since sooner or later all songs would be played anyways. Since they do not know that well other people's tastes rating is a good mechanism for preference elicitation.

In the second trial users described a context in which they would speak in advance about what type or genre of music they would agree to consume. Additionally they could decide on a common artist they like. Afterwards each of them would bring .mp3 files from the selected genre and listen with their friends in the same scenario. Since there wouldn't be the time to listen to all songs of one or several artists they are inclined to choose only top / popular songs of that artist making sure in advance that all group members agree with the group choice. In this way they would base their choices on the ratings given by large systems such as YouTube. One driver proposed that he would choose the music for the whole group since he is "responsible" for the trip. Others said that they would agree to bring music CDs with popular artists such as Lady Gaga.

5.5. GENERAL QUESTIONS

In the final stage of our discussions we asked users some more general questions.

5.5.1. OTHER SCENARIOS

First we wanted to get an insight of which other scenarios users would think about using GroupFun. They replied that they imagine longer trips or events of gathering of friends - e.g. birthday parties. Some users mentioned listening to music while working at the same place, practicing gym or for a football match. Others noted small or large social events.

5.5.2. FUTURE USE

When asked if they would use GroupFun in the future users confessed that they definitely will in the case they would spend more time on Facebook. Some emphasized on the need of better interfaces and transparent algorithms to be implemented.

5.5.3. SUGGESTIONS AND COMMETS

Finally we asked them if they have any suggestions or comments for us. All interviewees mentioned that the order of the playlist is very important and that they would want the system to play all the songs in the selected order. Furthermore, they agreed that as long as the songs are played, eventually, then they do not insist in voting on the position of the songs towards the first positions in the final playlist. Future suggested improvements need to focus on simplicity: drag and drop songs, synchronization with other devices, get ratings from listening frequency. User effort needs to be minimal.

Other proposed extensions focused on smart-phone apps: iPhone, Android, etc. Making the application available for people who are not necessary friends on Facebook may also be useful. Also GroupFun should use a simple algorithm. Some tag-based voting is a suggestion for improving the algorithm which should be easy to understand.

6. CONCLUSIONS AND FUTURE WORK

In our experiment we obtained valuable user feedback on algorithms performance from both subjective and objective measurements. We presented our results on the GroupFun user study regarding the 4 algorithms: PWS, DWS, LM and PS.

Our findings point to the fact that users take into account other members' preferences when initially eliciting individual preferences to some extent. They enjoy listening and discovering their friends tastes even though they had little knowledge in advance. Furthermore, some users consider other group members' preferences when they are reviewing recommended songs even though most of the times they start with a self-appreciation of the uploaded items. They use social aspects to determine individual satisfaction and align their preferences to the group. Out of the 4 algorithms they prefer PWS and DWS; the two algorithms have close scores. Both the deterministic and the non-deterministic versions of the weighted sum are excellent solutions for the implementation. The subjective evaluation is closely connected with the objective one. Users are more satisfied with the performance of PWS and DWS than LM and PS.

User feedback also highlighted the fact that some users like to be surprised by more random recommendations whereas some other times ratings should be equally considered. All participants agreed that GroupFun helped them to discover new music and were very satisfied with both their choices and the recommendation list. Based on the algorithm study we are able to derive several improvement guidelines for the GroupFun system: 1) minimize user effort by allowing individuals to easily manage their music from various devices 2) maximize group satisfaction through novel recommendations enhanced interaction and 3) consider each user equally important for a non-manipulable mechanism.

For the future we plan to extend our design for the GroupFun system. To learn more about the perceived ease of use and perceived usefulness of our application we plan to invite more members and analyze user feedback. We also intend to develop a new version of the algorithm which will match users' behavior better than the current one. This would also give us a platform for subsequent improvement of interaction and automation of negotiation techniques.

7. ACKNOWLEDGMENTS

We thank the participating users for their answers and involvement in the evaluation of our algorithms. Moreover, their feedback during discussion sessions was highly appreciated.

8. REFERENCES

1. Baur, D., Boring, S., Butz, A. Rush: Repeated Recommendations on Mobile Devices. Proc. IUI '10, ACM (2010), 91-100.
2. Hastie, R. and Kameda, T. The robust beauty of majority rules in group decisions. In Psychological Review (2005). Vol. 112, no. 2, 494-508
3. Masthoff, J. Modeling a Group of Television Viewers. Proc. Workshop future TV in Intelligent Tutoring Systems, ACM (2002), 34-42
4. Masthoff, J. Group modeling: Selecting a sequence of television items to suit a group of viewers. User Modeling and User-Adapted Interaction (2004), Vol. 14, no. 1, 37-85.
5. Masthoff, J. The pursuit of satisfaction: affective state in group recommender systems. In Computer Science (2005), Vol. 3538, 297-306
6. McCarthy J.F., Anagnost, T.D. MusicFX: an arbiter of group preferences for computer supported collaborative workouts. Proc. Computer Supported Cooperative Work, ACM (1998), 363-372
7. O'Connor, M., Cosley, D., Konstan, J.A., Riedl, J. 2001. PolyLens: A recommender system for groups of users. Proc. European Conference on Computer Supported Cooperative Work, ACM (2001)

APPENDIX A: ONLINE QUESTIONNAIRE

Dear Participant,

By participating in this survey, you help our research group understand how various algorithms and interfaces satisfy user needs. You and your friends can listen to your group music at anytime in GroupFun. Please note that your information will be kept confidential. Thank you very much for your participation. This survey was designed and distributed by the HCI GROUP, Swiss Federal Institute of Technology Lausanne (EPFL)

* Required
Top of Form

1. Background questions

1. Gender: *

- Male

- Female

2. Age group: *

- <20
- 20-24
- 24-29
- 30-34
- >34

3. Education *

- Highschool
- Bachelor
- Master
- PhD

4. In a typical week, I use social networks: *

- <5 hours
- 5-10 hours
- 10-15 hours
- >15 hours

5. I am familiar with music applications. * E.g. iTunes, Winamp, Windows Media Player, last.fm, Deezer.fr, etc.

	1	2	3	4	5	
Strongly disagree						Strongly agree

2. Background on relationship among users

1. I feel close to my friends. *

	1	2	3	4	5
Driver	○	○	○	○	○
Passenger1	○	○	○	○	○
Passenger2	○	○	○	○	○
Passenger3	○	○	○	○	○

2. I know my friends' musical tastes well. *

	1	2	3	4	5
Driver	○	○	○	○	○
Passenger1	○	○	○	○	○
Passenger2	○	○	○	○	○
Passenger3	○	○	○	○	○

3. Evaluation questions

1. I consider the most important factor(s) for a good playlist to be (check as many as you want): *

- ☐ Many of the songs are suggested by me
- ☐ The songs in the playlist are diverse
- ☐ The playlist helps me discover new music
- ☐ The playlist has a good transition between songs

2. I am satisfied with the songs recommended to me by: *

	strongly disagree = 1	2	3	4	5 = strongly agree
Alg1	○	○	○	○	○
Alg2	○	○	○	○	○
Alg3	○	○	○	○	○

	strongly disagree = 1	2	3	4	5 = strongly agree
Alg4	○	○	○	○	○

3. The system suggested an adequate number of songs. *

	1	2	3	4	5	
Strongly disagree	○	○	○	○	○	Strongly agree

4. The playlist is suitable for the whole group. *

	1	2	3	4	5	
Strongly disagree	○	○	○	○	○	Strongly agree

5. I consider the most important factor(s) for accepting a song which is not uploaded by me to be: *

- ☐ The song match my interest
- ☐ One of my friends likes the song
- ☐ Most of my friends like the song
- ☐ The song match the context

6. My friends' emotions influence mine. *

	1	2	3	4	5	
Strongly disagree	○	○	○	○	○	Strongly agree

7. GroupFun is useful. *

	1	2	3	4	5	
Strongly disagree	○	○	○	○	○	Strongly agree

8. I will use GroupFun again. *

	1	2	3	4	5	
Strongly disagree	○	○	○	○	○	Strongly agree

9. I will tell my friends about GroupFun. *

	1	2	3	4	5	
Strongly disagree	○	○	○	○	○	Strongly agree

10. I will listen to the group music even when I am alone. *

	1	2	3	4	5	
Strongly disagree	○	○	○	○	○	Strongly agree

APPENDIX B: INTERVIEW QUESTIONS

Part A: specific comments regarding the 4 outcomes

a) Would you like to change the final outcome provided by different algorithms? In which way? (Manipulation)
b) Do you consider the outcomes of the 4 algorithms to be fair? (Fairness)
c) Did you understand the inner logic of the system: how it recommends and computes top songs for the group? (Recommendation understanding)
d) Did you receive diverse recommendations? Where you familiar with them? How about enjoyable recommendations? Did they match your interest? (Diversity, enjoyable, match interest).
e) How would you prefer to discuss or negotiate about the outcomes.

Part C: general questions

a) Besides the current scenario, what others would you propose for this application?
b) Will you use GroupFun with friends in the future?
c) Do you have any suggestions/comments for GroupFun?