Breaking the filter-bubble

Using visualizations to encourage blind-spots exploration

Author:
Jayachithra Kumar

Thesis committee:
Prof.dr.ir. G.J. Houben (Chair)
Associate Prof.Dr. W.P. Brinkman
Assistant Prof.Dr. N. Tintarev

To obtain the degree of Master of Science in Computer Science
Data Science & Technology Track
To be defended publicly on August 31, 2018

Delft University of Technology
Faculty of Electrical Engineering, Mathematics & Computer Science
Department of Web Information Systems

August 23, 2018

Jayachithra Kumar: *Breaking the filter-bubble*, © August 31, 2018

STUDENT NUMBER:
4617312

SUPERVISOR:
Assist.Prof.Dr. N. Tintarev

Abstract

In recent years, personalized recommender systems have been facing criticism in research due to their ability to trap users in their circle of choices, called "filter-bubble", thereby limiting their exposure to novel content. In solving the issue of filter-bubble, past research has focused on providing explanations to users about *how* a recommender system recommends a specific item. This thesis addresses the issue of filter bubbles by helping users understand not just *why* a recommendation was made, but to also convey something about the limits of this recommendation.

In this thesis, we help users to better understand their consumption profiles by exposing them to their unexplored regions, thereby indirectly nudging them to diverse exploration. We refer to these unexplored regions as the user's *blind-spots*, and we visualize these by enabling comparisons between a user's consumption pattern with that of other users of the system. We compare the effectiveness of two visualizations – a bar-line chart and a scatterplot — in representing this consumption information and in increasing a user's intention to explore new content.

We performed a live user study to test our system (n=23). The results suggest that users are able to better understand their profile with both the visualizations. Furthermore, our results confirmed that users with a higher understanding of their profile tend to explore their blind-spot categories more. From our experiment, we provided a first step towards increasing user's awareness of their choices as well as providing the kind of user control that encourages users to explore new types of items.

Acknowledgements

I would first like to thank my thesis supervisor Dr. Nava Tintarev of Web Information Systems at TU Delft. The door to Nava's office was always open whenever I ran into a trouble spot or had questions about my research or writing. She consistently allowed this thesis to be my own work but steered me in the right direction whenever she thought I needed it.

I would also like to thank my fellow members of the Epsilon group for the stimulating discussions, insightful feedback and all the fun we had during the last one year. Special thanks to the participants in my survey, who have willingly shared their precious time during the process of interviewing. Without their passionate participation and input, the validation survey could not have been successfully conducted.

Finally, I must express my very profound gratitude to my family and friends for providing me with unfailing support and continuous encouragement throughout my years of study. Thank you for helping me through difficult times. This accomplishment would not have been possible without you.

Jayachithra Kumar,
August 2018

Contents

Abstract **v**

Acknowledgements **vii**

1 Introduction **1**

1.1 Motivation 1

1.2 Research Goal 2

1.3 Results 3

1.4 Contribution 3

1.5 Thesis Outline 4

2 Background and Related Work **5**

2.1 Breaking the Filter bubble 5

2.2 Approaches 6

2.2.1 Algorithmic approaches 7

2.2.1.1 Diversity approaches 8

2.2.1.2 Serendipity approaches 9

2.2.1.3 Novelty approaches 10

2.2.1.4 Coverage approaches 11

2.2.2 User-centric approaches 12

2.2.2.1 One-shot approaches 12

2.2.2.2 Dynamic approaches 14

2.2.3 Discussion 16

2.3 Evaluation 17

2.3.1 User-centric Evaluation 18

2.3.1.1 Multi-criteria online studies 18

2.3.1.2 Targeted online studies 20

2.3.2 Discussion 23

2.4 Conclusion 24

2.5 Future work and Gaps 24

2.5.1 Gaps 24

2.5.2 Future Work 25

3 Identification and Visualization of Music Consumption Pattern **27**

3.1 Introduction 27

3.2 Application Setting 27

3.3 Feature selection 28

3.4 Music Data Extraction 29

3.4.1 Global Profile Extraction 30

3.4.1.1 Genre extraction 31

3.4.1.2 Artist hotness extraction 34
3.4.1.3 Consumption data extraction 35
3.4.2 User Profile Extraction 36
3.4.2.1 Genre extraction 37
3.4.2.2 Artist Hotness extraction 38
3.4.2.3 Consumption data extraction 38
3.5 Frequent genre pattern extraction 39
3.5.1 Recursive Elimination Algorithm (RElim) 39
3.5.2 Implementation 40
3.6 Choice of Visualization 42
3.6.1 Visualization 1: Scatterplot 43
3.6.2 Visualization 2: Bar-line chart (Baseline) 44

4 Evaluation 1 - to study user's understanding of visualization 49
4.1 Introduction 49
4.2 Design 50
4.3 Independent Variable 50
4.4 Dependent Variables 50
4.5 Hypotheses 51
4.6 Procedure 51
4.7 Materials 53
4.8 Participants 54
4.9 Measures 55
4.10 Results 56
4.11 Discussion 61
4.12 Limitations 64

5 Evaluation 2 - to study user's music exploration 67
5.1 Introduction 67
5.2 Design 67
5.3 Independent variable 67
5.4 Dependent Variable 68
5.5 Hypothesis 68
5.6 Procedure 68
5.7 Materials 70
5.8 Participants 70
5.9 Measures 71
5.10 Results 71
5.11 Discussion 71
5.12 Additional Observations 73
5.13 Limitations 74

6 Conclusion 77
6.1 RQ1: Are visualizations effective in conveying to users, their consumption pattern and blind-spots? 77
6.2 RQ2: Which visualization is the most effective in representing to users, their consumption patterns and blind-spots? 78

6.3 RQ3: Does user's understanding of their profile correlate with their intention to explore their blind-spots? 78
6.4 Limitations 79
6.5 Future work 80
6.6 Concluding remarks 80

A Counter-balancing the order of visualizations and questionnaires 83
A.1 Order of Visualization 83
A.2 Order of questionnaires 83

Bibliography 85

List of Figures

2.1 Approaches to combat filter-bubble 7
2.2 Balancer system 13
2.3 2-D scatterplot - social recommendation visualization 14
2.4 User-controlled filtering in social network 15

3.1 Extraction and visualization of consumption pattern 28
3.2 MSD - songs per year 31
3.3 Data integration for global profile. 32
3.4 Distribution of Genres 33
3.5 Genre distributions in All Music and Tagtraum datasets 34
3.6 Artist hotness distribution in Million song dataset 35
3.7 Distribution of play counts 36
3.8 Distribution of tracks across users 36
3.9 Pre-pocessing steps for RElim algorithm 40
3.10 RElim procedure 41
3.11 Example scatterplot visualization used in the study 45
3.12 Example bar chart visualization used in Movie Lens 46
3.13 Example bar-line chart visualization used in the study 47

4.1 Conceptual model for Stage 1 50
4.2 User instructions - scatterplot 53
4.3 Global and individual consumption pattern in scatterplot 54
4.4 Global and individual consumption pattern in bar-line chart 54
4.5 Participant demographics by gender and age 55
4.6 Participant demographics by music listening frequency 55
4.7 Global consumption pattern in scatterplot 64

5.1 Conceptual model for Stage 2 68
5.2 Exploration phase 69
5.3 Spotify's genre mixing interface 70
5.4 Exploration interface used in the study 70

List of Tables

2.1 Diversity approaches . . . 9
2.2 Serendipity approaches . . . 10
2.3 Novelty approaches . . . 11
2.4 Coverage approaches . . . 12
2.5 One-shot approaches . . . 14
2.6 Dynamic approaches . . . 16
2.7 Multi-criteria online studies . . . 20
2.8 Targeted online studies . . . 23

3.1 List of all MSD features . . . 29
3.2 Available Music Dataset comparison . . . 31
3.3 Genre distribution of 'All Music Genres' dataset . . . 32
3.4 Comparison of existing music APIs . . . 37
3.5 Frequency values for top 20 frequent item-sets (globally) . . . 42
3.6 Average artist hotness for top 20 frequent item-sets (globally) . . . 42
3.7 Scatterplot attributes mapping . . . 44
3.8 Bar-line chart attributes mapping . . . 46

4.1 Demographic questionnaire . . . 51
4.2 Participant demographics - Profession . . . 55
4.3 Participant demographics - Musical background . . . 56
4.4 Participant demographics - Music search frequency . . . 56
4.5 Mean (std) scores of participants for identification of their first and second most consumed genre. . . . 57
4.6 Mean (std) scores of participants for identification of artist hotness values of their first and second most consumed genre. . . . 57
4.7 Mean (std) confidence scores of participants for identification of their first and second most consumed genre. . . . 58
4.8 Mean (std) confidence scores of participants for identification of the artist hotness of their first and second most consumed genre. . . . 59
4.9 Mean (std) scores of participants for identification of their first and second highest blind-spot genres. . . . 59
4.10 Mean (std) scores of participants for identification of artist hotness of their first and second highest blind-spot genres. . . . 60
4.11 Mean (std) confidence scores of participants for identification of first and second highest blind-spots. . . . 60
4.12 Mean (std) confidence scores of participants for identification of artist hotness of their first and second highest blind-spots. . . . 61
4.13 Post hoc user feedback analysis - bar-line chart . . . 63

4.14 Post hoc user feedback analysis - scatterplot 63

5.1 Post-stage assessment questionnaire 69

5.2 Correlation between exploration factor and user's actual understanding of their profile. 72

5.3 Correlation between exploration factor and user's perceived understanding of their profile. 73

5.4 Total number of new genre/genre-combinations explored by the user. . . 74

6.1 Mean(std) user rating for perceived aspects of visualization 78

A.1 Questionnaire split 84

A.2 Distribution of users for each question set - visualization combination . . 84

1 Introduction

1.1 Motivation

In this thesis, we aim to increase user's awareness of their *filter-bubble* in *Recommender Systems* by exposing them to their un-explored regions, thereby indirectly nudging them to diverse exploration. We refer to these un-explored regions as a user's *blind-spots*, and we intend to visualize their blind-spots by enabling comparisons between a user's consumption pattern with that of other users of the system.

Recommender Systems (RS) are software tools that provide suggestions for items that are most likely of use to users. Depending upon the use-cases of recommender systems, the recommendations made by these systems could vary, aiding in various decision-making processes ranging from what music to listen to or what movies to watch, to what products to buy or what news to read. Personalized RSs work by trying to predict what the most suitable products or services are to a user based on factors such as user's preferences, past consumption, demographics, context, etc. By considering these factors, such a personalized system would aid in effectively narrowing down the myriad available options to few meaningful choices that the users are most likely to consume.

In recent years, personalized RS has increasingly helped businesses in retaining customers and increasing sales revenues. The RS of Netflix accounts for up to 75% of what users watch and has helped to save $1 Billion each year [33]. Spotify's RS has helped in increasing its number of monthly users from 75 million to 100 million [1]. More and more users are relying on these RSs for their daily news consumption, e-commerce, etc., and with or without their knowledge, they are also feeding information about their preferences to the system, thereby letting the system shape its future recommendations.

In practice, Personalized RSs work by continuously studying user's behavior in order to build a unique profile for each user of the system [27]. Based on this profile, the RS then filters, prioritizes and hides information from the users. This results in situations where, for two different users, different results get shown for the same query. In extreme cases, users are even prone to receive biased information that confirms to their viewpoint as opposed to unbiased information. In other words, users are likely to get trapped in their own circle of choices - called their "filter-bubble". The effects of filter-bubble are manyfold, and depending upon the domain of RS its magnitude may vary. For example, in the case of News RSs, users might not be able to see opposing viewpoints in their

[1] https://www.bloomberg.com/news/articles/2016-09-21/spotify-is-perfecting-the-art-of-the-playlist - accessed June 2018

news feed, which in turn creates strong and polarized views/opinions among users. In case of Music and Movie RSs, user's exposure to novel content would be limited, and their exploration would be hindered. In general, irrespective of the domain of RS, filter-bubble tends to show an adverse effect on user's freedom of choice and autonomy, and in the long run, on the quality of consumed information.

Throughout literature, several solutions have been proposed to combat the issue of filter-bubble, and one of the most common and well-explored solutions include introducing those items in the recommendation list, that are outside of a user's regular consumption spectrum. Accordingly, diverse items are recommended without notifying the users, or explaining why the item was recommended (e.g., [85, 72, 42, 74]). More emphasis was later placed on user-centric aspects of RSs. Transparency of the underlying recommendation algorithm was increased by providing explanations as to *why* a particular item was recommended (e.g., [24, 51, 76, 68]). Some systems attempted at giving full control to users to alter their filter-bubble (e.g., [61, 77, 48, 79]). But this does not guarantee unbiased exposure since, if the user wants to stay in their filter-bubble or shrink their filter-bubble, the controls would let them do that. Therefore, while all the above approaches help broaden a user's information exposure and make them understand the underlying system, only a small focus has been given to increase user's awareness or inform them about the limitations of the system. So it is time that we take a step back and focus on the two core issues - *breaking the filter-bubble* whilst *preserving user's autonomy*.

1.2 Research Goal

In this thesis, we focus on giving users a holistic view of their filter-bubble - by enabling them to compare *their* consumption pattern with the (aggregate) consumption pattern of other users of the system ('global' consumption pattern). By providing such a high-level information, we intend to increase user's awareness of their filter-bubble and to verify if such an awareness motivates users to explore items outside their filter-bubble, without getting lost in the plethora of options.

The primary goals of our research are two-fold. Firstly, we would like to lay the groundworks for a system as discussed above by analyzing some of the parameters that could make the system most effective. More specifically, we examine the *means* to convey such high-level information, and its impact on increasing user's awareness of their filter-bubble and encouraging exploration. To such a degree, visual interfaces have been recently gaining attention in recommender systems, and as some recent studies suggest (e.g., [78, 61]), using visualizations enable better content discovery in RSs compared to traditional methods of content display such as ranked list interfaces. We compare two types of visualizations - bar-line chart and scatterplot - and evaluate the effectiveness of these visualizations at conveying to users, their music consumption pattern and their blind-spots (regions that are under-consumed).

Secondly, we would also like to study whether such an increase in user's awareness (if any) has an effect on user's intention to explore new content. Such a correlation would bolster the significance of increasing user's awareness, and eventually, by studying user's exploration, it would enable us to distinguish between user's lack of interest and lack of awareness about the un-explored content.

These two goals bring us to our main research question:

> ***How effective are visualizations, at increasing user's awareness of their filter-bubble, and in turn, increasing user's inclination to explore content outside their filter-bubble?***

This research question raises the following three sub-questions:

1. Are visualizations effective in conveying to users, their consumption pattern and blind-spots?
2. Which visualization is the most effective in representing to users, their consumption blind-spots?
3. Does user's understanding of their profile correlate with their intention to explore their blind-spot genres, and if it does, what is the strength of such a correlation?

1.3 Results

In this thesis, we introduce a method to increase user's awareness of their unexplored regions - *consumption blind-spots* - by visualizing their consumption pattern, and providing a means to compare it with the consumption pattern of other users of the system. We compare the performance of two visualizations - *bar-line chart* and *scatterplot* - in effectively conveying user's consumption information. Furthermore, we extend this study to elucidate if, such an increase in user's awareness of their blind-spots has a positive effect on their intention to explore their un-explored content. Such an effect would bolster the significance of increasing user's awareness and autonomy in recommender systems.

We conducted a user-centric evaluation of our system, to compare both bar-line chart and scatterplot representations of the user's consumption profile obtained from their Spotify accounts. We tested the user's understanding of each of these visualizations, and our results show that both the visualizations increase user's awareness about their consumption pattern and blind-spots.

We further analyzed the user's exploration pattern, and found a significant positive correlation between the user's awareness and their exploration of their blind-spot genres. This shows that, given the autonomy and awareness, users are indeed willing to expand their bubble. Additionally, our results show that users appreciate an opportunity to listen to tracks by combining genres, and they prefer an interface that lets them combine their frequent and blind-spot genres.

1.4 Contribution

In our thesis, we study the effects of using visualizations in increasing user's awareness of their profile, and in turn nudging diverse exploration. Our experiment demonstrates that bar-line chart and scatterplot visualizations are capable of conveying to users, information about their consumption pattern and blind-spots, thereby, increasing their awareness of their blind-spots. Future research should focus on other types of visualizations and interfaces to convey such information, and identify the one that is best suited for the purpose.

Furthermore, we identified that users with higher understanding of their consumption pattern explore their blind-spot genres more. Such a correlation strengthens the significance of focusing on user-centric aspects of recommender systems and keeping users in the loop during the recommendation process.

Additionally, we also discovered that, while exploring music using our exploration interface, users were most interested in exploring genre combinations that combined their frequent and blind-spot genres. This observation is especially significant since, using our system we provide sufficient control to users to expand their preferences, without the risk of exacerbating their filter-bubble.

1.5 Thesis Outline

Chapter 2: We discuss work related to the topics addressed in this thesis. In particular, we discuss the different algorithmic and user-centric approaches introduced in literature to combat the issue of filter-bubble. We also discuss different online and offline evaluation metrics used to test the effectiveness of these approaches. We provide an extensive analysis on the specific areas where these approaches are inadequate in relation to our work, highlighting the gaps and positioning our thesis in terms of these existing works.

Chapter 3: We dive into our proposed approach that overcomes some of the shortcomings of existing methods. We explain, in detail, the main steps involved in the extraction of different parameters necessary for the identification and visualization of blind-spots. For each parameter, we discuss the design decisions and implementation methodology.

Chapter 4: We perform an online evaluation of both our visualizations (bar-line chart and scatterplot) to study their effects on increasing user's understanding about their profile.

Chapter 5: We extend our first evaluation to further observe user's exploration pattern, and study if it is correlated with user's understanding of their profile.

Chapter 6: We return to our three research questions as stated in the beginning of this chapter, which address the effectiveness of visualizations and their impact on user's exploration, and answer them with respect to the results obtained through chapters 4 and 5, and through our post-hoc analysis.

2

Background and Related Work

Article selection procedure

The selection of article took place through Google Scholar using keywords: filter-bubble, beyond-accuracy, recommender systems, online evaluation, user-centric approaches, serendipity, diversity, novelty, and coverage. Only the most recent articles and those with the most relevant contributions were selected for this literature review. Around 70% of the papers covered here were published in the last eight years (2010-2017). Papers from 2010 through 2014 were chosen if they were cited at least ten times and papers from 2015 through 2017 were chosen if they had a citation count of at least five. A total of 65 papers were considered to be relevant for the study. These papers covered both journals and conference papers with 47% of the articles published in ACM (Association for Computing Machinery) conferences. After the initial analysis these works were classified in the following four sub-categories (two under approaches category and two under evaluation category): algorithmic approaches (24 articles), user-centric approaches (12 articles), offline evaluation (20 articles) and online evaluation (11 articles), with 6 articles occurring in both algorithmic approaches and offline evaluation.

2.1 Breaking the Filter bubble

Recommender systems aim to provide item suggestions that are most likely of interest to users. By providing a personalized recommendation, recommendation algorithms become further effective by improving the quality of recommendations and by overcoming the information overload problem. However, an overt focus on personalization might be dangerous in many ways. Firstly, the recommended items might get more and more similar to existing items thereby failing to consider human desires for variety and diversity, and eventually leading to boredom. In extreme cases, this might also lead to the recommendation of misrepresented facts and spurious information that is tailored to bias people. Secondly, too much personalization to a user's taste might hamper a user's personal growth and experience by trapping the user in his/her small circle of choices, called "filter-bubble" [27]. In some cases, this could lead to limiting the user's exposure to information from like-minded people, creating the so-called "echo chambers" [15]. In the long run, getting trapped in these filter bubbles tend to subject users to develop potentially extreme and polarized views.

Throughout literature, several solutions have been proposed to combat the problem of filter bubble. Focusing too much on improving the ’accuracy’ of recommendation algorithms and viewing algorithmic accuracy as synonymous with ’usefulness’ of recommender system was seen as a major setback in breaking the filter bubble. Soon there was a shift in focus, from improving the accuracy of recommendations, to incorporating more diverse and novel recommendations, withstanding slight compromises in accuracy. New ’beyond-accuracy’ objectives like serendipity, diversity, novelty, and coverage were introduced, and recommender systems incorporating these objectives were found to be more satisfying and useful to the users than the traditional recommender systems. Furthermore, implementing these beyond-accuracy objectives also meant that new metrics are needed to evaluate the system, as the traditional accuracy metrics won’t apply anymore. Hence equal focus was given in research for finding new evaluation metrics for each of the beyond-accuracy objectives, and finding methods to combine different objectives and define proper trade-offs between them.

Given all the algorithmic methods of coping with filter bubble, none of these would be of much use if users themselves are not aware of their filter bubble. Diverse recommendations, in some cases, might even confuse users and decrease their trust in the system. Hence, for a recommendation system to be accepted by users it is important that users understand the system, as in, *why* a particular recommendation was made. Keeping this in mind, several user-centric implementations of recommender systems were made lately, with its main focus on increasing user’s understanding of the system through visualizations and explanations. Eventually, recommender systems are built to serve users, and therefore, user-centric approaches to evaluation of recommender systems are deemed crucial as they help in understanding the recommender system with respect to user’s subjective experiences.

In this section, we will discuss several approaches proposed to combat the issue of filter bubble and the different evaluation metrics available in the literature to measure the effectiveness of these approaches. The rest of the literature review is organized as followed: Approaches to combat filter-bubble are discussed in Section 2.2, with algorithmic approaches in Section 2.2.1 and user-centric approaches in Section 2.2.2. A brief discussion on the approaches is provided in Section 2.2.3. The user-centric evaluation metrics are discussed in Section 2.3, followed by discussion in Section 2.3.2. Finally, the chapter is concluded in Section 2.4

2.2 Approaches

Methods to combat filter-bubble in RS research were initially focused in news recommendation systems and later spread to other domains like music and movies. Based on the primary objectives of these works, it is evident that a general solution to bursting the filter-bubble should involve the following two inter-dependent steps:

1. **Exposing** people to contradictory information
2. Nudging people to **consume/react to** contradictory information

Exposing people to contradictory information can be achieved either directly using algorithmic approaches or indirectly using user-centric approaches. In algorithmic approaches, exposure is provided by *directly* including 'beyond-accurate' items in the recommendation list. These include items which are different from the items in a user's regular consumption spectrum. User-centric approaches mainly focus on increasing user's awareness of the filter bubble thereby *indirectly* motivating them to seek more diverse information. The algorithmic approaches operate in the *production stage* with a focus on *avoiding* the creation of filter-bubble whereas the user-centric approaches operate in the *post-production stage* where the main aim is to *deal* with the already created bias of the produced content.

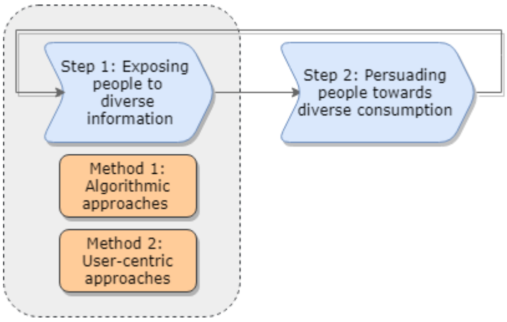

Step 1: Exposing people to diverse information
Step 2: Persuading people towards diverse consumption
Method 1: Algorithmic approaches
Method 2: User-centric approaches

FIGURE 2.1: Approaches to combat filter-bubble

Once users are exposed to information outside their filter bubble using either of the methods, then in step two, it is made sure that users actually understood and consumed the contradictory information. This is achieved in the literature by obtaining feedbacks from users either directly using questionnaires or indirectly by observing their interaction with the system [54, 55, 29]. These user feedbacks can then be used to refine the type of information preference (in step one) for each user, and hence the inter-dependence. In literature, however, there is much less work in step two compared to the works in the first step. This could be attributed to the level of risk associated with attempts to nudge people to consume diverse information and interact with the system. In extreme cases, users could get annoyed with the system and eventually lose trust in the system. The approaches in the second step are out of scope for my study, and in this survey, I focus on the different approaches in the first step.

2.2.1 Algorithmic approaches

Algorithmic approaches work by tuning the recommendation algorithms to recommend items that are *different* from the regularly consumed items. This difference can be defined in several ways and based on how they are defined, four different beyond-accuracy concepts exist - diversity, serendipity, novelty, and coverage. Throughout literature, several definitions exist for each of these concepts, and in some cases, these definitions tend to overlap. However, a few widely accepted definitions mark a clear distinction between these concepts. For example, an item is - diverse, if it is different (mostly content-wise) from the other items; serendipitous, if the item is both attractive and surprising to the user; and novel, if the item is previously unseen by the user.

Coverage is defined on a system level, and a recommendation list is said to have good coverage if the list covers the catalog of all available items.

Diversity is generally approached as a simple list re-ranking problem, and in some cases, diversity-optimized algorithms are developed especially with recent advances in matrix factorization methods. Graph-based approaches are most widely used for implementing serendipity and novelty, with more focus on long-tail (rare) items in novelty. Approaches to achieve coverage also comprise mainly of graph theoretic approaches with a focus on long-tail items. In this section, we focus on various approaches to implementing these beyond-accuracy metrics.

2.2.1.1 Diversity approaches

The notion of diversity was originally introduced in information retrieval research where the list of documents with diverse information and broader topic coverage was found to increase user satisfaction [21]. In Recommender systems, diversity is studied from either a user-centric perspective where diversity is defined as the internal differences in items in recommendation list or from a system-centric perspective where diversity is defined as the overall differences between different recommendation lists. User-centric diversity is also known as individual diversity or intra-list diversity and system-centric diversity as aggregate diversity or inter-user diversity [32].

Several approaches have been proposed to increase the diversity of a recommendation list, and these approaches belong to one of the two categories - *Recommendation Re-ranking* for diversity or *Diversity Modeling*.

Recommendation Re-ranking: As its name suggests, in re-ranking approaches the results generated by the existing recommendation algorithm are re-ranked in order to produce a shorter, more diverse list, while maintaining the relevance of the list. Most widely used re-ranking approach is the greedy re-ranking method where the objective function is given as a linear combination of the relevance of the item and the average distance to items in the list. A slightly different adaptation of this method was used by Smyth and McClave (2001) [73] in case-based recommender systems, where the relevance of the item indicates the similarity between the query and a case and distance is computed as the complement of similarity. Ziegler et al. [6] proposed a re-ranking strategy to balance top-N recommendation lists according to the user's full range of interest. In this approach, the relevance was defined as the item's relevance given by the collaborative filtering algorithm, and the distance was measured from the taxonomy-based similarity metric. Apart from the greedy re-ranking approach, several advanced approaches like Maximal Marginal Relevance (MMR) technique [21] are used in recent literature, which focuses on the relationship between an item's relevance and its similarity to the other items in the list. Vargas et al., (2011) [84] adopted two objective functions from search diversity - IA-select and MMR scheme, where, IA-select model tradeoffs between relevance and diversity to minimize user's dissatisfaction. In 2014, Vargas et al., [83] proposed a method by which "binomial diversity" was used to measure the genre diversity by using a binomial distribution. The objective function is a combination of binomial diversity and relevance of the item. Quadratic optimization methods were also adopted to increase diversity. For example, Zhang and Hurley (2008) [89] reduced the problem of diversity as a joint maximization problem of two objective functions representing an adequate level of similarity and item diversity.

Diversity Modeling: One of the main advantages of using a re-ranking approach is their ease of deployment to existing systems. This is because in re-ranking, the underlying recommendation algorithm is treated as a black-box and re-ranking is applied to the recommendations generated by the algorithm. Furthermore, with re-ranking, the level of diversification can be tuned explicitly. Recent research focuses on "diversity-optimized" recommendations where new recommendation algorithms are used to generate diverse recommendations. Matrix factorization is one of the most commonly used recommendation techniques and most of the methods in this section use this technique in their implementation.

Two of the most common adaptations include using Portfolio Theory of Information Retrieval and Pairwise learning to re-rank approaches. The portfolio theory of information retrieval proposed by Wang and Zhu (2009) [85] quantifies a ranked list of documents with maximum relevance and minimum variance. This theory was later extended by Shi et al. (2012) [72] where the authors used portfolio theory to relate the level of diversification in the recommendation list with the user's taste obtained from his/her ratings. In the second adaptation, Pairwise learning to re-rank, user and item factors are learned by minimizing an optimization function that defines the difference between the original and predicted ranking for item pairs. Hurley (2013) [42] used this concept to produce a diversity-aware version for datasets with implicit user feedback with item dissimilarity used to define the objective function. Finally, Su et al. (2013) [74] proposed a diversification model based on pairwise learning to rank where the model operates on a set of items rather than individual items in the list. Similarity between item sets is computed pairwise as the product of item pair's latent factor vectors and diversity is defined as the aggregate similarity of all item pairs in the set.

TABLE 2.1: Diversity approaches

Authors	**Re-ranking based**	**Diversity modeling**
Smyth et al. [73]	X	
Ziegler et al. [6]	X	
Yu et al. [87]	X	
Carbonell et al. [21]	X	
Vargas et al. [84, 83]	X	
Zhang and Hurley [89]	X	
Ribeiro et al. [70]	X	
Wang and Zhu [85]		X
Shi et al. [72]		X
Hurley [42]		X
Su et al. [74]		X

2.2.1.2 Serendipity approaches

The term "serendipity" was first defined by Van Andel [13] as the "process of finding valuable and pleasant things that are not looked for". This term was later adapted in

IR and RS literature by Ge et al., [31] who defined serendipitous items as items that are both "attractive" and "surprising" to the users. The attractiveness of an item is generally attributed to its relevance, and hence the definition of serendipity mostly relies on the definition of "surprise".

Attempts to increase serendipity was first introduced in IR literature where software agents were built for discovering serendipitous information during web crawling[20]. In RS literature, serendipity was first introduced in content-based recommender systems [56], where serendipitous heuristics is used to introduce surprising recommendations. A three-step process is used to recommend items with a supervised learning module that classifies items as interesting to the user or not according to user preferences. Items for which the prediction was uncertain were considered serendipitous and hence kept in the list.

Graph-based approaches were also quite widely used to increase serendipity. For instance, Onuma et al. (2009) [50] translated the problem of discovering serendipitous item to a node selection problem on a graph, where nodes that are well connected to older choices, as well as unrelated choices, are given a higher score. This bridging score is then combined with item relevance score to provide the final recommendations. Nakatsuji et al. (2010) [7] proposed a graph-based approach where user graphs are created with weighted edges denoting the similarity between the users. Random walks with Restarts (RWR) are then performed in the graph to find users who are connected but not too similar, and surprising recommendations are obtained from these users. Zhang et al., [12] proposed a hybrid music recommender system which combines Latent Dirichlet Allocation (LDA) with listener diversity. The LDA model builds latent clusters of users, and artists are represented as a distribution over these clusters. Artists that are outside of a user's cluster are then assumed to provide serendipitous items. Finally, Adamoupolos and Tuzhilin (2014) [1] presented an approach to recommend serendipitous items based on the item's *unexpectedness* which is measured from the item's distance from a set of expected items. Recommendations are then provided by a utility function by combining an item's relevance score and its unexpectedness.

TABLE 2.2: Serendipity approaches

Authors	Preference based	Graph based	Cluster based	Distance based
Iaquinta et al. [56]	X			
Onuma et al. [50]		X		
Nakatsuji et al. [7]		X		
Zhang et al. [12]			X	
Adamoupolos et al. [1]				X

2.2.1.3 Novelty approaches

A novel recommendation item is defined as an item that is previously unknown to the user. Novelty is closely related to serendipity in the sense that, an item that is serendipitous must be both novel and surprising but an item that is novel is not necessarily serendipitous [36]. Novelty was first introduced in IR studies by Baeza-Yates et al., in 2009 [14] where a retrieved item was assumed to be novel if it is both relevant and unknown to the user. This concept was extended in RS literature to define novel items as items that are either unknown to the user or different from what the user has seen. Kapoor et al. (2015) [52] later identified that novel items could also include items that are known to the user but forgotten. In recent studies, the term novelty has been attributed to the popularity of an item such that, the most popular an item is, the most likely it is to be known by the user.

Most of the attempts at increasing novelty in recommender systems are based on this last definition where rare/less popular items or "long-tail items" are recommended as novel items to the user. For example, Ishikawa et al., (2008) [43] used diffusion theory to deal with the long tail phenomenon in a collaborative filtering based recommender system. Recommendations were made by observing the trend with which new information spreads among users, such that, users who first discover an item are identified as the "source" of novel recommendations. Zhou et al. [11] proposed a method that uses a user-item graph to increase novelty, which works by first assigning weights to the items that are rated by a given user followed by equal distribution of the weights of each item to other users. This was later extended by Liu et al. (2012) [45] where the novelty of the recommendations was further improved by assigning higher weights to users with a fewer number of ratings. Another graph-based approach was proposed by Shi (2013) [71] who defined a cost flow model based on first order Markovian graph with transition probabilities between user-item pairs in which the cost to reach the target user node from an item node is computed by propagating the cost scores through the edges. Items with lower costs are then given higher priority during recommendation.

TABLE 2.3: Novelty approaches

Authors	Diffusion theory	Graph based
Ishikawa et al. [43]	X	
Zhou et al. [11]		X
Liu et al. [45]		X
Shi [71]		X

2.2.1.4 Coverage approaches

Coverage is defined as the degree to which the items in the recommendations cover the catalog of all the available items [36]. In literature, two types of coverage have been identified namely - "user coverage" which represents the extent to which all users are covered by the system and "item coverage" which represents the extend to which the system covers all the available items. Item coverage is further classified, by Herlocker et al. [36], into "prediction coverage" which represents the ratio of items in the item catalog for which predictions can be made by the system and "catalog coverage" which represents the ratio of items that actually appear in the user's recommendation list.

Similar to novelty, most of the works on increasing coverage are based on long-tail items or rare items in the recommendation list. One of the first works that explicitly focused on increasing coverage was proposed by Adomavicius and Kwon (2011) [3] in which a graph-theoretic approach was used to maximize recommendation coverage in movie recommendation system. Users and items are represented as vertices of the graph, and an edge exists between a user and an item if the item is predicted as relevant to the user. Top N items with high coverage are then predicted by solving the maximum flow problem which gives the largest possible number of recommendations that can be made from all the available items such that each user is provided with a maximum of N recommendations. The second approach proposed by Adomavicius and Kwon in 2012 [2], followed a ranking based technique to promote long-tail items in the recommendation list. The approach works by setting a threshold parameter on the popularity of the item, and the items that are above the threshold are recommended. This method allows for explicit trade-off control between accuracy and coverage. Finally, Vargas and Castells (2014) [82] proposed a greedy optimization technique to increase genre coverage in collaborative filtering recommender systems. The approach works by first constructing a neighborhood for the item and then selecting all the items for which the target item appears in their neighborhood.

TABLE 2.4: Coverage approaches

Authors	**Ranking-based**	**Graph based**	**Nearest neighbor based**
Adomavicius et al. (2011)[3]		X	
Adomavicius et al. (2012)[2]	X		
Vargas and Castells [82]			X

2.2.2 User-centric approaches

User-centric approaches combat filter-bubble, by using as their medium, those aspects of recommendation systems that are directly visible to users - such as placement of items in the list, the interaction of the interface and visualization. These approaches indirectly contribute to combating the filter-bubble by creating awareness about the already existing bubble to the users. All approaches in this section fall into one of the two categories - *one-shot* approaches, where users are exposed to information (in the form of visualization or notifications) that makes them aware of their filter bubble, and *dynamic* approaches - where users are given a level of control (in the form of interactive interfaces) over the parameters that influence their filter bubble. Most of the early works in user-centric approaches were mainly one-shot and aimed at news recommendation systems to increase diverse political exposure.

2.2.2.1 One-shot approaches

Among the different design choices available in RS research, one of the most commonly used approaches to deal with extreme polarization involves exposing people to information that are usually *hidden* from them. Items might be hidden as a result of filtering mechanism and personalization, and these hidden regions of the recommender systems are generally termed as *blind-spots* [62]. Xing et al. (2015) [86] exposed information hidden by search engines by developing a browser widget called 'Bobble' that lets users compare their Google search results with the results of other users. Each time a user issues a search query, Bobble captures the query and reissues it from a large number of vantage points. This way, in addition to their own results, users can also see the results that are hidden from them, thereby making them aware of their personalization. This method reveals hidden information to the level of individual items. But this may not be feasible in recommender systems where there is a large number of items and most of these items usually remain hidden from the users.

To deal with this issue, Tintarev et al., [62] introduced two modifications: representing items at genre-level rather than individual level, and using visualizations, rather than plain text, as the medium to represent the gaps in user's profiles. Specifically, visualizations were used to display to users their most frequent genres and the global consumption pattern, with different color codes providing the distinction. Experiments were conducted with both bar charts and chord diagrams and the results showed that chord diagrams were more effective in explaining to users about gaps in their consumption. However, the experiment was conducted with selected top genres and whether the visualization will have the same effectiveness and readability with more genres is unknown.

In recommender systems, as the system gains more and more information about user's preferences their recommendations become either more narrowed down or more broadened over time. Providing users with an overview of their consumption pattern might help them see this behavior and motivate users to seek for more diverse information. Munson et al., in 2013 [60] used this logic to nudge people to consume balanced political viewpoints. The authors developed a browser widget called 'Balancer' (Figure 2.2) which visually displays the left-right balance of the news articles viewed by the user with the help of a stick figure trying to balance on a rope. The political inclination of a user is represented by an imbalance in the figure thereby informing users of their reading bias. Initial analysis of the experiment suggested a noticeable change in reading behavior, with users going towards a more balanced exposure after seeing the feedback as compared to a control group. Another browser widget, called Scoopinion[1] generalizes this concept by providing a visual summary of one's reading habits by tracking all the news sites and stories read by the user so far. The site also provides recommendations based on user's reading habits but the system has not been tested in an academic setting and hence the effectiveness of the system is remains unknown.

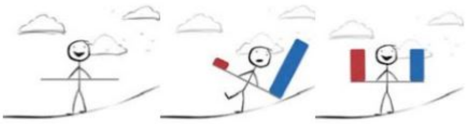


FIGURE 2.2: Balancer extension displays bias in a user's news reading preferences [60]

As an alternative to the above-mentioned category of approaches, few works exist that try to increase user's exposure to diverse content by leveraging interfaces that expose people to content beyond the classic ranked list. For example, Tsai et al [78] proposed using a two-dimensional scatterplot interface (Figure 2.3) to display the results of recommendation algorithm for a social recommendation function of the Conference Navigator System [17]. The effect of the scatterplot interface on the *exploration* of social recommendations by academic attendees was evaluated and ranked list interface was used as the baseline. A between-subjects study was conducted, and the study combined both

[1] www.scoopinion.com, accessed January 2018

subjective and objective system evaluation. Results showed that the visual interface encourages users to explore a more diverse set of recommended items and the exploration pattern of the visual interface was more extensive, covering items from all categories, not just top-ranked items. The authors later extended this study in [80], and this time users were allowed to tune the features/dimensions of the scatterplot visualization in order to include four dimensions - academic feature, social feature, interest feature and distance feature. A within-subjects user study was conducted to evaluate the system where the performance of dual interface with the scatterplot and ranked list was compared against the baseline interface using just ranked list interface. The results of the study show that the scatterplot interface helps users explore more diverse items compared to ranked list interface especially when more than two dimensions are involved. Furthermore, Users of scatterplot interface showed significantly higher coverage measurements between tasks.

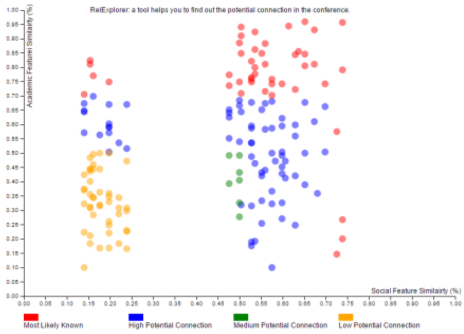


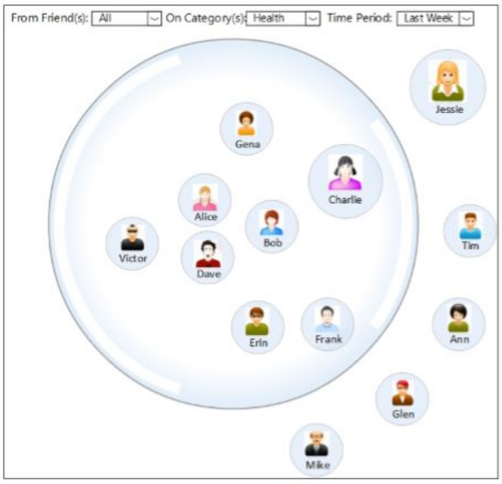
FIGURE 2.3: Two-dimensional scatterplot visualization proposed by Tsai et al. to display results of conference navigator system. Red - most likely known, Blue - high potential connection, Green - medium potential connection, Yellow - low potential connection [78]

TABLE 2.5: One-shot approaches

Authors	Exposing blind-spots	Exposing consumption pattern
Xing et al. [86]	X	
Tintarev et al. [62]	X	
Munson et al. [60]		X
Scoopinion		X
Tsai et al. [78]	X	
Tsai et al. [17]	X	

2.2.2.2 Dynamic approaches

Most of the works under this section focus on effectively utilizing interactive interfaces to expose people to diverse information. Along that line, one of the most common design choices to achieve this is using *user-controlled filtering* techniques to support content discovery in recommender systems. For example, Nagulendra and Vassileva [61] developed a visualization and control tool in 2014 to display to users their filter bubble in their social network (Figure 2.4). The visualization is provided by representing a huge bubble with friends and topics inside the bubble influencing the user's news feed and the ones that are outside having less/no impact. Users can control of their filter bubble by dragging and dropping the items inside and outside their bubble. This tool not only provides users with an idea about which other users and topics influence their news feed but also maximizes the user's control over their filter bubble.



From Friend(s): All
On Category(s): Health
Time Period: Last Week

FIGURE 2.4: Nagulendra and Vassileva's software allows users to control their filter-bubble [61]

In 2015, Tintarev et al., [77] extended this concept for content discovery in micro-blogs, specifically Twitter. Filtering was provided on aspects that influence news feed like communities, networks structure and ranking of tweets, and user control was provided with the help of an interactive interface where users could provide hop-values for each community to stress its influence. For example, a community with 0-hops will be excluded from tweets, a community with 1-hop will include tweets from people who follow the community and a community with 2-hops will additionally include tweets from the followers of the people in the community. By changing the hop values, users could either expand or contract their filter bubble. Experimental results showed that participants found the system useful for discovering new content and exploring community structure.

Another similar content discovery tool, called 'HopTopics', was developed in 2016 by Kang et al. [48] for Twitter. The system combats the effect of filter bubble by extending the traditional Twitter feed to include social connections (friends, friends-of-friends, etc.) up-to n-hops. This helps users to gain access to information sources beyond a user's typical information horizon. User experiments show that the interface and interaction model were effective for improving the perceived transparency and control compared to baseline interfaces.

In 2017, a tool called RelExplorer was developed by Tsai et al., [79] which supports user-controlled content discovery in a system recommending co-attendees in an academic conference. Similar to the above systems, the main aim of this system is to increase the control and transparency of recommender systems by having a human-in-loop system. Recommendation is provided by three recommender systems each of which recommends co-attendees to a conference based on their academic feature (similarity of past publications), social feature (social network distance) and interest feature (similarity of interests). Users can assign weights to each feature by adjusting their corresponding sliders that range from 0-100. The system was evaluated with the help of user study at an academic conference, and the results showed that the system is perceived useful for users to find social contacts at conferences.

All the above-mentioned content discovery tools aim to break the filter bubble indirectly by providing control to users on the content that would influence their feed, all while keeping the underlying recommendation algorithm as a black box. While providing users with controls might enhance content exploration and improve the transparency of recommendation algorithms, it does not guarantee a change in their filter-bubble. A user who is unaware of the effects might even choose to shrink their circle of influence thereby exacerbating their filter-bubble. Keeping this in mind, few works in the literature focus on using interactive interfaces to increase user's awareness by showing *hidden* information, just as discussed in section 2.2.2.1.

Siamak Faridani et al. (2010) [29] developed an online tool called 'Opinionspace' which helps users to visualize and navigate through diverse comments in online news articles, videos, blogs, etc. In traditional systems, most of the times users are shown only the top comments (i.e., comments with most likes, for example) and less popular comments are hidden due to space constraints. This issue is handled here by applying dimensionality reduction and projecting the data as an arrangement of points on a two-dimensional plane, with farther the point (comment) the more diverse it is. Visualization is interactive, in that, when readers hover over a particular point, the comment related to the point is shown, and they are prompted to rate the comment and give their level of agreement with the comment. This encourages users to deliberate on multiple viewpoints and avoids user's bias towards only reading popular comments.

Rbutr[2] is a Chrome add-on that informs users when a web page is disputed, rebutted or contradicted elsewhere on the Internet. When a user visits a rebutted page, the add-on informs the user about the rebuttal and displays the rebutting articles. Users could also add opposing viewpoints for an article or search for rebuttals for an article. The main aim of Rbutr is to avoid misinformation and promote critical thinking by showing hidden information. The effectiveness of the system, however, is quite unclear since it is not tested in an academic setting.

TABLE 2.6: Dynamic approaches

Authors	Exposing blind-spots	User-controlled filtering
Nagulendra and Vassileva [61]		X
Tintarev et al. [77]		X
Kang et al. [48]		X
Tsai et al. [79]		X
Siamak Faridani et al. [29]	X	
Rbutr	X	

2.2.3 Discussion

The creation of filter-bubble is a slow process that happens over time as a result of filtering and over-personalization. Even though all the techniques discussed so far provide a quick solution to this issue, whether these techniques are persistent over time is not

[2] http://rbutr.com/

known. Periodic tests need to be conducted over a long term to judge the true effectiveness of the system, and currently, none of the existing works acknowledge/address this issue. One main reason could be the high cost incurred in conducting such long-term experiments which makes it infeasible especially when dealing with user-centric evaluation.

In RS research one of the most highly preferred user-centric approaches involves using explanation interfaces that explain to users *why* a particular recommendation was made. Explanation interfaces have shown to be useful in increasing the transparency of recommender systems and helping users accept new recommendations [76]. For example, in RelExplorer [79] four explanation components (social similarity, co-authorship, publications, text similarity) are provided to justify the recommendation of a co-attendee to an academic conference. These explanations are mainly aimed at enhancing the user's trust with the system. However, it remains an open research area when it comes to works relating explanations with filter-bubble.

Finally, when it comes to the user-centric approaches discussed in this paper, I consider works that *indirectly* motivate users to seek diverse information by increasing their awareness and without explicitly nudging them. This is important because prompting users to consume diverse information would be of no use when users themselves are not aware of their filter-bubble. Furthermore, there is a level of risk associated with explicitly nudging users to consume information since it could be easily misperceived as information imposition. That being said, few approaches still exist, mainly in the domain of politics, that explicitly prompt users to consume and react to diverse information [54, 55]. These approaches fall under a separate category (step two of figure 2.1) and are out of scope for my study.

2.3 Evaluation

In traditional recommender systems, algorithmic accuracy was considered as the most important determinant of usefulness of the system. Therefore, in most of the traditional systems, it was common practice to use offline accuracy (similarity) metrics such as precision, recall and F-1 measure to evaluate the system. Soon with the advent of beyond-accuracy objectives, accuracy metrics became obsolete, and new distance-based and graph-based metrics were introduced [83]. These metrics provide a good estimation of system-centric beyond-accuracy aspects from an algorithmic perspective. But the results do not necessarily translate to usefulness and satisfaction for the user. Therefore new user-centric (online) evaluation metrics were introduced in order to validate the offline results and to make sure the offline scores match user-perceived values. Online studies are also considered essential to evaluate user-centric approaches with interface-level implementations such as visualizations, interactions, controls, etc. which rely on human cognitive ability. Few studies combine both offline and online metrics - where the former is used to measure algorithmic aspects of the system while the latter is used to measure interface level aspects such as usefulness of explanations (explanation interfaces) [76], organization of recommendation lists [38], positioning of items in the list [32], etc. - to obtain best results.

In recommender systems research, extensive studies have been made on offline metrics. But these are out of scope for our thesis, and in this section, we will discuss the different online evaluation metrics and frameworks introduced in the literature.

2.3.1 User-centric Evaluation

In order to better understand the performance of recommender systems from the user's perspective, recent literature focuses on user-centric implementation and evaluation of recommender systems. One of the earliest applications using user-centric evaluation of recommender system focused on evaluating user's trust in recommender system [66]. This section provides an overview of works that rely on subjective evaluation of recommender systems.

Typically, all user-centric evaluation metrics use one of the two models - *between-subjects* design or *within-subjects* design. The main difference between both these models lies in the number of independent variables used in the experiment. In a between-subjects design, users are provided with one algorithm at a time whereas, in a within-subjects design, users compare multiple algorithms. In general, the between-subjects design provides a more realistic view of the use of recommender system whereas within-subjects design can be used to evaluate and compare multiple algorithms. In addition to these two categories, Kaminskas et al. (2016) [47] provided a more specific classification where they categorize the research works based on the beyond accuracy aspects measured in the evaluation. Accordingly, the first category contains all works where the *relationships* between different user perceived recommendation qualities are measured, and the second category contains works that study the effect of *specific* algorithms or user-interface level adaptations on *specific* beyond-accuracy objectives. This classification is more relevant, and hence in this section, I extend this model to also include aspects other than beyond accuracy objectives (like user satisfaction, confidence, etc.). For ease of representation, the categories are re-named as 'Multi-criteria online studies' and 'Targeted online studies' respectively without affecting the meaning.

2.3.1.1 Multi-criteria online studies

One of the first works based on multi-criteria user studies was performed by Pu et al. (2011) in their paper "A user-centric evaluation framework for recommender systems" [67]. They provide an extensive evaluation framework called ResQue which aimed at explaining how the perceived quality of the recommendation influences user's beliefs, attitude towards the system and satisfaction with the system, and how these factors indeed influence user's behavioral intentions. They measure the perceived quality of the system with the help of two beyond accuracy objectives - novelty and diversity. Due to the meticulous distinction between novelty and serendipity, the latter was ignored in the study as it might confuse users. The framework consisted of an extensive set of 31 questions grouped into 15 categories. The experiment was performed with 239 participants of diverse nationalities where users were first asked to select a product of their choice from an online site, and then fill out the answers to the evaluation questionnaire from the framework. The results showed that the perceived usefulness of the system was considerably influenced by perceived novelty and moderately by the perceived diversity.

Knijnenburg et al. [2012] [5] proposed an extensive framework for evaluating the user experience of recommender systems from objective system aspects by using subjective system aspects as mediators. The framework is based on a set of six structurally related concepts namely objective system aspects **(OSA)** (such as recommendation algorithm and presentation), subjective system aspects **(SSA)** (such as perceived quality and diversity), perceived experience **(EXP)** (i.e., attitude), interaction **(INT)** (i.e., behavior), situational characteristics **(SC)** (such as privacy concerns) and personal characteristics **(PC)** (such as trust). Several experiments were conducted to study the effects of one or more of these variables in the framework. User experience is measured from concepts such as perceived accuracy, diversity, system effectiveness, choice difficulty, etc., by defining a chain of effects to draw relationships between these aspects. The results show an inconsistent relationship between the actual and perceived diversity between the three algorithms used in the study (i.e., k-nearest neighbor, matrix factorization and generally most popular algorithm). For example, the perceived diversity of k-NN algorithm with no actual diversity is higher than the perceived diversity of the same algorithm with a little actual diversity. However, this condition does not hold for the other two algorithms. Similarly, an increase in perceived diversity tends to increase the perceived accuracy of the system thereby increasing the overall user experience. Finally, in case of "generally most popular" algorithm, diversification of the algorithm is shown to be as effective as replacing it with a recommendation algorithm.

In "User perception of differences in recommender algorithms" [9], Ekstrand et al. compared three collaborative filtering algorithms - item-item CF, user-user CF, and singular value decomposition (SVD) CF, on five dimensions - novelty, diversity, accuracy, satisfaction, and personalization. The framework of Knijnenburg et al., [5] was used to model the evaluation and MovieLens user community was recruited as participants. A within-subjects study was conducted and each user was assigned to two out of the three algorithms. In the first step, users were provided with two lists of movies with ten recommendations side-by-side on the same interface, and they were asked to choose their most preferred list based on the first impression. Secondly, users were asked to answer a set of 22 questions about various aspects of the list, and finally, users were asked to choose their most preferred algorithm. The results show that the perceived values of diversity and novelty correlate with the measured values, and that, diversity has a positive influence on user's choice of the system and novelty has a significant negative impact on user's satisfaction. As a result, the user-user CF algorithm, which has the highest novelty among the three algorithms, was least preferred compared to the other two algorithms.

Finally, Fazeli et al. [2017] [10] in their work, try to find the relation between user-satisfaction of a recommender system measured using online evaluation and the accuracy of the system measured offline. User satisfaction is evaluated based on five quality metrics such as perceived usefulness, accuracy, novelty, diversity and serendipity which was taken from the ResQue framework. The experiment used a between-subjects design and involved users of Open Discovery Space (ODS) which is an e-Learning environment. Three best algorithms were chosen from the offline evaluation, and these were used for online evaluation. These algorithms are a memory-based CF (User KNN), a graph-based CF and a model-based CF(matrix factorization). Each user evaluated recommendations from a randomly assigned algorithm, and each algorithm was assigned to 20 users. The questionnaire consisted of a set of six questions that measure the five quality metrics. Results reveal that, even though traditional evaluation suggests User KNN as the best algorithm, users were satisfied with the accuracy of all the algorithms regardless of the choice of the algorithm. The authors suggest that there is no point in finding the most accurate algorithm if users do not recognize the differences between recommender systems.

TABLE 2.7: Multi-criteria online studies (All the aspects measured are user-perceived and categorized based on Knijnenberg's framework [5])

Authors	SSA	EXP	INT	SC	PC
Pu et al. [67]	X	X	X		X
Knijnenburg et al. [5]	X	X	X	X	X
Ekstrand et al. [9]	X	X			
Fazeli et al. [10]	X	X			

Using the framework suggested in Knijnenburg et al. [5] as a baseline, comparisons between the above works were made (table 2.7). The impact of personal and situational characteristics on system aspects remain one of the most under-studied concepts in these studies. However, from the results of experiments in [5], it is evident that these characteristics do tend to influence the overall experience of the user to a great extent. For example, users with higher domain expertise tend to have higher perceived diversity compared to users with lower expertise. Understanding such relationships may help in accurately personalizing recommendations by effectively tailoring to specific users.

Furthermore, out of the four beyond accuracy objectives, only novelty and diversity are mostly addressed in these studies. Serendipity was ignored in Pu et. al [67] in order to avoid confusing users. One reason for not measuring the concept of coverage could be because it has multiple dimensions and that it is defined at system level and not user level.

2.3.1.2 Targeted online studies

Recent literature provides several targeted user studies which were conducted in order to evaluate a specific subjective systems aspect (SSA). These studies can be classified based on the specific beyond accuracy objectives they measure and the different types of independent variables they use in the study.

Diversity vs Perceived Usefulness

Most of the research focuses on measuring the effect of diversity on user's perception of system usefulness by comparing different recommendation algorithms as their independent variable.

Along these lines, one of the earliest works on user-evaluation of recommender systems were done by Ziegler et al. (2005) [6], where a topic-diversification algorithm was proposed in order to increase the diversity of recommendation lists. An online evaluation was done with more than 2,100 users with data from BookCrossing.com and Amazon.com. The effects of diversity were measured on two collaborative filtering (CF) algorithms, and a between-subjects study was conducted. Users were first displayed with a list of 10 recommendations from one of the two algorithms and asked to rate the items in the list after which they were presented with a set of questions that measure the user's perceived diversity and satisfaction with the recommendations. The results showed that an increase in diversification tends to increase overall user satisfaction, which was notably significant in case of item-item CF algorithm.

A similar evaluation approach was followed by Castagnos et al. (2013) [22] to measure the impact of diversity on user-satisfaction in three different algorithms - CF, CB and Popularity-based filtering (POP). The experiment was conducted in the movie domain, and the study involved four steps. Each participant was first asked to rate movies provided by one of the three algorithms assigned to him/her. Users were then provided with three recommendation lists from the three algorithms and were asked to order the list based on their preferences. Finally, a post-questionnaire was presented to measure the perceived relevance, diversity and confidence level of the recommendation. The results show that diversity has a positive influence on user's satisfaction. However, too much diversity tend to confuse users and have a negative impact if a proper explanation is not provided.

While measuring the effect of diversity between different algorithms, the above-mentioned approaches assume the same level of diversity for all the algorithms, and for all the users. For a single algorithm, different levels of diversity might have different effects on user satisfaction. Keeping this in mind, Willemsen et al. (2011) [8] conducted a within-subject user study to evaluate how different levels of diversification affects the perceived diversity of the list. The main aim of the study is to prove that increasing the diversity of a recommendation list helps in effectively decreasing the choice-overload problem. The experiment was conducted on a movie recommendation system that uses matrix factorization algorithm. The study consisted of 97 participants, and each participant was asked to rate three different lists of varying diversity. Questionnaires were used to measure perceived diversity, perceived attractiveness, expertise, choice difficulty and trade-off difficulty. The evaluation framework was based on [5], and the results showed that increasing the diversity of the list tend to increase the perceived diversity and decrease the choice and trade-off difficulty.

Interface Types vs Perceived Diversity

Few other studies evaluate the influence of different interface-level aspects on user's perceived diversity. For example, Hu et al. [38] studied how an organization interface compares to a standard list interface in terms of perceived usefulness and diversity. In an organization interface, the recommendations are categorized based on some common trade-off properties (e.g., products which are "cheaper but heavier" than the chosen product are grouped together). The experiment was conducted on a commercial perfume website, and a total of 20 participants were recruited for the study. A within-subjects study was conducted, and the evaluation framework was based on ResQue from [67]. All participants used both the interfaces assigned to them in random order and answered a set of questions about their general preferences, usefulness, informativeness, etc. The results were then analyzed to measure two main aspects of RS - *categorical diversity* (difference among categories) and *item-item diversity* (difference among items). The results showed that users perceived categorical diversity much stronger than item-item diversity and that they perceived organization interface to be much more useful than list interface.

Another aspect of RS interface that influences user's perception of diversity is the placement of items in the list. Ge et al. (2011) [32] conducted a study to find the most effective placement of diverse items in the recommendation list and the extent to which diversity influences the perceived quality of the recommendation system. Static lists of movies were created for three genres with three different placements of diverse items. A within-subjects pilot study was conducted with ten users, and each user was provided with all the three lists. Users were asked to rate movies and answer four questions related to the user's satisfaction, diversity, and surprise. Results showed that diverse items in the list aroused user's interest and attention and that discovery of diverse items were much more effective when they were arranged in a block than when they were distributed across the list. The results also showed that users were more interested to read additional information about diverse items, thereby stressing the importance of providing explanation interfaces in recommender systems.

Novelty vs Perceived Quality

Apart from diversity, other aspects like Novelty and Serendipity were briefly evaluated in few studies. The experiment by Celma in 2009 [23] was aimed at measuring the Novelty of the recommendation systems by explicitly asking users if they were familiar with the item or not. A within-subjects study was conducted in order to compare the novelty in three algorithms - CF, content-based audio similarity (CB) and hybrid approach. The study was performed on last.fm users and consisted of several rounds. In each round, participants were asked to rate songs from a list of 10 recommended songs and the rating was collected based on familiarity - whether the user knows the song, and quality - whether the user likes the song. Novelty is considered to be the inverse of familiarity, and the results show that the list with higher novelty was perceived to be of lower quality compared to the list with higher familiarity. Adding meta-data and explanation for why a particular song was recommended was proposed as a possible solution to improve user's acceptance of novel recommendations.

Serendipity vs Perceived Usefulness

Finally, Zhang et al. (2012) [12] conducted a user-study on their serendipitous framework called *Auralist* in order to study the impact of serendipity, novelty, and diversity on user satisfaction. They developed three hybrid versions of their basic music recommender system - *Community-Aware Auralist* aims to provide diversity, *Bubble-Aware Auralist* recommends artists outside user's music-bubble and *Full Auralist* is a hybrid of the first and second algorithm. User study involved 21 participants and it was aimed at comparing the perceived serendipity, novelty, enjoyment and overall qualitative satisfaction between *Basic* and *Full Auralist*. Each user was presented with 20 recommendation from both the recommendation algorithms in random order and for each algorithm questions were asked to measure these different aspects. Results showed that users found *Full Auralist* to be more useful than *Basic Auralist* even though there is a slight compromise in accuracy in the former algorithm. Serendipity and novelty are seen as a positive contributor to user's satisfaction.

TABLE 2.8: Targeted online studies

Authors	SSA	EXP	INT	SC	PC
Ziegler et al. [6]	X	X			
Celma [23]	X	X			
Willemsen et al. [8]	X	X			X
Hu et al. [38]	X	X			
Ge et al. [32]	X				
Zhang et al. [12]	X	X			
Castagnos et al. [22]	X	X			

Out of all the system-specific aspects addressed in the above works, serendipity, and coverage remain under-explored compared to diversity and novelty. However, on further analysis, it is evident that the two main aspects of serendipity, i.e., "surprise" and "usefulness" are measured independently in several studies. For example, in Ge et al. (2011) [32], the authors measure how surprising the recommendations are to the users and in Hu et al. (2011) [38] and Zhang et al. (2012) [12], the authors measure the perceived usefulness of the recommender systems. Therefore by including questions that correlate "surprise" and "usefulness" of a list we could possibly measure the perceived serendipity of the recommendation system.

2.3.2 Discussion

On studying both offline and online metrics, one could clearly notice an imbalance in the number of works in both the areas. Despite the apparent need for user-centric evaluations in recommender systems, most of the systems rely on just offline metrics and only a few works in this area actually include user-centric evaluations. The reason behind this could be justified by the high cost of conducting such user-centric evaluations, especially in designing the experiment - like setting the right questionnaires, finding the appropriate number of participants, etc. Moreover, user-centric evaluations intend to measure subjective system aspects, and hence a careful attention to details - such as choosing the order of displaying results, measuring personal user aspects (like expertise), etc. - needs to be made in order to avoid biased results.

Another reason why online metrics are not widely preferred could be attributed to the inconsistencies in results. The same experiment conducted using different evaluation frameworks or with different questions might yield opposite results, and factors such as user's demographics might contribute to the bias. For example, while discussing multi-criteria online studies (section 2.3.1.1), the results obtained by Pu et al., [67] showed that an increase in novelty increases user satisfaction while the study by Ekstrand et al.[9] provided contradicting results. One reason for this could be attributed to the differences in question formation. While in [67], the users were directly asked to answer*"The items in the list are novel"*, in [9], novelty was measured from indirect questions like *"has movies that you do not expect?"*. Furthermore, the nature of the question *"has more pleasantly surprising movies?"* in [9] correlates to 'serendipity' aspect, but it has been associated with 'novelty' which might have introduced a certain bias in the result.

Finally, out of the seven works discussed under 'Targeted user studies' (Section 2.3.1.2), four works insist the importance of providing explanations and meta-data in shaping user's perception of diversity. These works operate in three different domains - [22] and [32] in movie domain, [23] in music domain and [38] in e-commerce - and hence it can be interpreted that explanation interfaces play a crucial role in increasing the perceived usefulness of recommender systems, irrespective of the application domain.

2.4 Conclusion

In this survey, we discussed the different approaches that attempt to mitigate the effect of filter-bubble in recommender systems, and various evaluation metrics that test the effectiveness of these approaches. Although throughout literature, most of the focus has been given to algorithmic approaches, the importance of user-centric approaches has been widely acknowledged. Several works have been discussed which prove the effectiveness of user-centric approaches in increasing user's awareness about their filter bubble. These approaches function by exposing users to hidden information, and by providing users with an overview of their consumption patterns.

When it comes to evaluation metrics, it has been acknowledged that any evaluation of recommender system in terms of beyond-accuracy metrics is not reliable if it does not involve user feedback. The importance of user-centric evaluation lies in its ability to reflect the quality impressions perceived by the users in a much better way than offline metrics. Recent literature has shown a shift in focus from using just offline evaluation to using online evaluation or ideally both. This is especially evident in scenarios where one must decide between two algorithms with similar offline scores. One main drawback of online evaluation lies in the inconsistency of the evaluation framework used throughout literature. Following a uniform framework would prevent inconsistent results, and this is one area that is lacking in the literature.

2.5 Future work and Gaps

Below I summarize the gaps that have been identified from the literature survey.

2.5.1 Gaps

- Few approaches have been developed that consider user-centric aspects of implementation
- Very few approaches focus on increasing user's awareness of their filter-bubble

- No standard/uniform online evaluation framework is followed throughout the literature
- Few approaches use visualizations to enable content discovery in recommender systems
- Effect of explanation interfaces in breaking the filter-bubble is under-explored

2.5.2 Future Work

Based on some of the gaps identified in literature, future work aims to achieve the following:

- To identify the effect of visualizations in increasing user′s awareness of their filter-bubble. More specifically,
  - To identify which visualization provides a better understanding of a user′s consumption pattern to the users.
- To set a starting point to be able to differentiate between a user′s lack of awareness and lack of interest in the content that is un-explored. More specifically,
  - To understand if a user′s awareness of their profile motivates exploration of new information by the user.

3

Identification and Visualization of Music Consumption Pattern

3.1 Introduction

The growing impact of filter-bubble implies that, in order to provide an unbiased recommender system, it is required to keep users in-the-loop during implementation and evaluation of the system. While several approaches focus on providing diverse recommendations to users and increasing the transparency of recommender systems, none of these approaches try to convey the limitations of these recommender systems, nor do they focus on increasing user's awareness of their filter-bubble.

The goal of this thesis is to alleviate some of the drawbacks of existing systems by increasing user's awareness of their potential filter-bubble by visualizing genre gaps ('Blind-spots') in their music consumption pattern. Consequently, we also aim at studying the change in their preferences (if any), given their awareness in their filter-bubble.

In this chapter, the concrete setting within which the thesis is placed is discussed, followed by the design decisions and considerations underlying the creation of visualizations. We first provide the setting of our thesis and a brief overview of the stages involved in the extraction and presentation of music genre blind-spots in Section 3.2. This is followed by a detailed discussion on each aspect of the procedure. Music feature elicitation and data elicitation are discussed in Sections 3.3 and 3.4 respectively. A detailed description of the extraction of music consumption pattern is explained in Section 3.5, and the choice of visualizations are explained in Section 3.6.

3.2 Application Setting

In this thesis, we focus on giving users a holistic view of their filter-bubble - by enabling them to compare *their* consumption pattern (user profile) with the (aggregate) consumption pattern of other users of the system ('global' consumption pattern or 'global' profile). In case of Music RS, multiple sources can be combined to obtain this consumption data. Music data sources like 'Million Song Dataset' [16] and 'Spotify' [53] offer several additional music features like genre, song loudness, artist, album information, etc., which provides a reasonable ground for analysis. Hence we resort to building and testing our hypotheses in the music domain of RS.

When it comes to visualizations, we do not aim to explain individual items to users, but rather highlight the important aspects of their profile as a whole (i.e., by grouping tracks based on genres). This way visualization could scale better and still provide an accurate representation of global and user's preferences. However, in doing so, we do not fail to consider the differences within a single category between the user and global profiles (section 3.3). Besides, in addition to representing a range of categories, we also aim to represent the interaction between these categories (i.e., when a track belongs to more than one genre). This enables us to highlight the specific categories users are most familiar with, thereby increasing user's trust in the visualization.

Figure 3.1 provides a brief overview of the stages involved in the extraction and visualization of consumption pattern. Steps 1 & 2 involve feature extraction and data collection respectively. Step 3 involves extracting global and local preferences using frequent item-set mining algorithm. Once the global and local preferences are extracted, visualizations are constructed to represent this data (step 4). The following sections describe in detail, the design decisions that went into each of these stages.

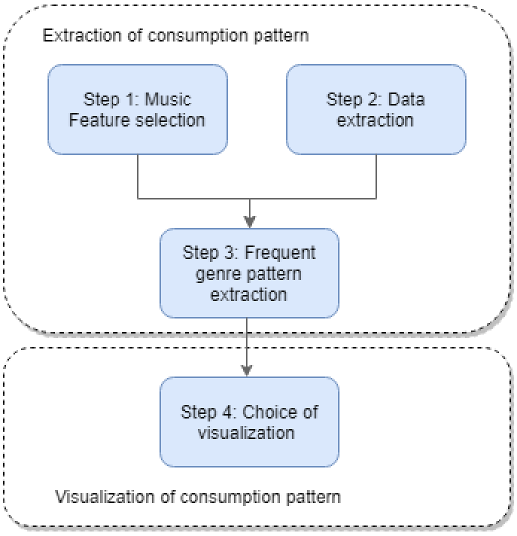

Extraction of consumption pattern
Step 1: Music Feature selection
Step 2: Data extraction
Step 3: Frequent genre pattern extraction
Step 4: Choice of visualization
Visualization of consumption pattern

FIGURE 3.1: Steps involved in the extraction and visualization of consumption pattern

3.3 Feature selection

We aim to represent a user's music blind-spots indirectly by comparing their music consumption pattern with the global consumption pattern. However, in doing so, we expect that the consumption data is not only well representative of their preferences but also well scalable to be represented in the form of visualization. To achieve this trade-off, we group individual items in such a way that their characteristics are still preserved on a high-level. Genre-based grouping provides a good collective representation of a user's music behavior, and it is a feature that users can easily relate to since it has already been used in existing recommender systems like Spotify. Hence we use *genre* as our first feature.

Furthermore, while representing tracks in genre-based groupings, we grant flexibility by allowing interactions between these genres. This is especially significant in scenarios where an item does not strictly belong to a single category. For example, a track could belong to both 'Rock' and 'Pop', in which case it will be represented by the category 'Rock, Pop' to indicate the interaction between both genres. Such a combination of genres helps in accentuating the specific categories that the users are most familiar with, and the categories that users consume in combination with other categories.

In addition to providing such an interaction between categories, it is also important for the system to be able to distinguish between a user's profile and global profile for the same category. Under the same genre, two user's might have different preferences in terms of artists, albums, tempo or even time-periods. The system should be able to highlight some, if not all, of these differences in the visualization. In order to achieve this, a second dimension is added to the visualization. To select the most representative feature we looked into the Million Song Dataset (MSD) which provides a total of 55 features for each track (Table 3.1). Out of these features, acoustic features might seem to provide an insightful representation of the tracks. But for users who lack sufficient musical background, these features would prove meaningless. Therefore we analyzed the remaining artist and song features, and chose **artist hotness** as the most representative feature. 'Artist hotness' (represented as 'artist_hotttness' in MSD) is a value (0 to 1), assigned by MSD for each artist, which corresponds to how much buzz the artist is getting right now. This value is computed algorithmically based on information derived from several sources, including mentions in the web, mentions in music blogs, music reviews, play counts, etc. [1]. This feature was chosen for the following reasons:

1. In comparison to other features, artist hotness is proven to provide a stable representation of user's preferences [40] .
2. It provides a good estimate of the user's taste equally for new and experienced users of the system [40].
3. Compared to other acoustic features, it is easier to explain to users and to understand.

TABLE 3.1: All music features available in MSD

Category	**Features**
Acoustic features	bars, beats, sections, segment, loudness, pitches, timbre, tatums, key, mode, tempo
Song metadata	song id, track id, sample rate, energy, duration, release, danceability, song hotness, title, year
Artist metadata	artist id, terms, similar artists, artist hotness, artist familiarity, artist location, artist latitude, artist longitude, artist name, musicbrainz tags

3.4 Music Data Extraction

In our visualization, we aim to compare the global consumption pattern (global profile) with the consumption pattern of individual users of the system (user profile), as a means

[1] https://musicmachinery.com/tag/hotttnesss/, as of March 2018

to represent their blind-spots. More specifically, for each user, we aim to visualize their most preferred genres (and genre-combinations) and average artist-hotness value for each genre. To achieve this, ideally we need such a dataset that could provide us with the following information:

1. **Globally most preferred genres and genre-combinations**. This entails the most frequently consumed genre/genre-combinations across all users of the system.
2. **Aggregated Artist Hotness value for each of the above genres**. This says where most of the user's preferences lie in the spectrum of least to most popular artists.
3. **Specific user's most preferred genre and genre combination**. This data is computed for each user of the system based on his/her listening history, and it is used to draw comparisons of a user's genre preferences with the global preferences.
4. **Aggregated Artist Hotness value for each of the above genres**. This information says where the specific user's preferences lie in the spectrum of least to most popular artists.

Currently, none of the existing data sources provide the above information for global and user profiles. However, there are individual data sources that provide different aspects of the required information. For example, the 'EchoNest' API [28] provides a global profile containing triplets with **user ID**s of several users of an undisclosed system, the **track ID**s of the tracks they listened to, and the **playcounts** of each track. On the other hand, MSD dataset provides the artist hotness value for each of these tracks. Therefore, by merging, pre-processing, filtering and aggregating these and other data sources, we can obtain all the necessary information. In this section, we delineate the design decisions that went into the selection and pre-processing of several data sources to obtain the final data for visualization. Steps are described in detail for both global and user profiles separately. For each of these profiles, we describe how we extract consumption data, genre data and artist hotness information.

3.4.1 Global Profile Extraction

To build the global profile, we need the global listening pattern (i.e., tracks listened by all users), genre/genre-combinations, and artist hotness values of these tracks. Throughout literature, several datasets have been provided for research on music recommender systems, and each of these datasets entails different features. A brief overview of the available datasets and information provided by them can be found in the Table 3.2.

Among the available datasets, Million Song Dataset (MSD) by Bertin-Mahieux et al. [16] is the largest dataset, containing audio features, song and artist meta-data for a million contemporary music tracks. The core of the dataset contains music features and meta-data extracted using EchoNest API [28] (now acquired by Spotify). MSD dataset was scraped in 2011, and hence it has tracks from the years 1922 - 2011. Figure 3.2 shows the distribution of tracks. MSD is one of the most widely used and readily available datasets in research for Music RSs. Moreover, data required to match the track IDs in MSD with track IDs from other data sources such as Spotify are readily available. Hence, citing the above reasons, we chose MSD as the main data-source for extraction

TABLE 3.2: Available Music Dataset comparison

Datasets	MSD [16]	Last.fm [75]	Yahoo! Music [26]
Number of Songs	**1,000,000**	5,05,216	1,36,000
Unique Artists	44,745	2,94,015	97,812
Number of triplets (user-item ratings)	48,373,586	17,559,530	7,17,000,000
Number of users	10,19,318	3,59,347	18,00,000
Type of rating	play count	play count	Rating
Year	1922-2011	1922-2011	2002-2006

of the global profile, and we merged the other missing information, such as genre tags, into this dataset.

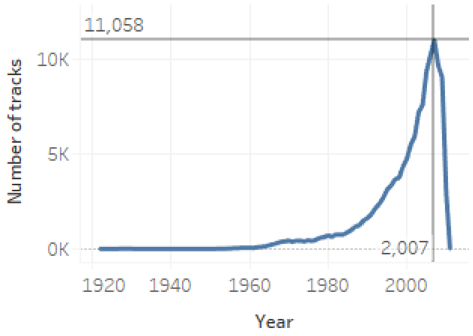


FIGURE 3.2: MSD - songs per year

In the following sections, we discuss various databases involved in building the global profile and how they were processed and merged. A general outline of the data integration for the global profile is shown in Figure 3.3. Here the datasets in yellow are the source datasets, the ones in white are intermediate data sets, and the one in green is the final dataset for the global profile.

3.4.1.1 Genre extraction

One main drawback of MSD is that it does not provide genre information associated with tracks. To deal with this issue, in music RS research genre information is usually obtained from either acoustic features [81, 58] or annotation of tags [64] using genre classification tasks. Directly obtaining genre information from music features is out of scope for our study, and hence we resorted to existing research to use their already extracted genre information [34, 41]. We merged genre information from these datasets with the individual tracks from MSD based on their *track_ids* and *song_ids*.

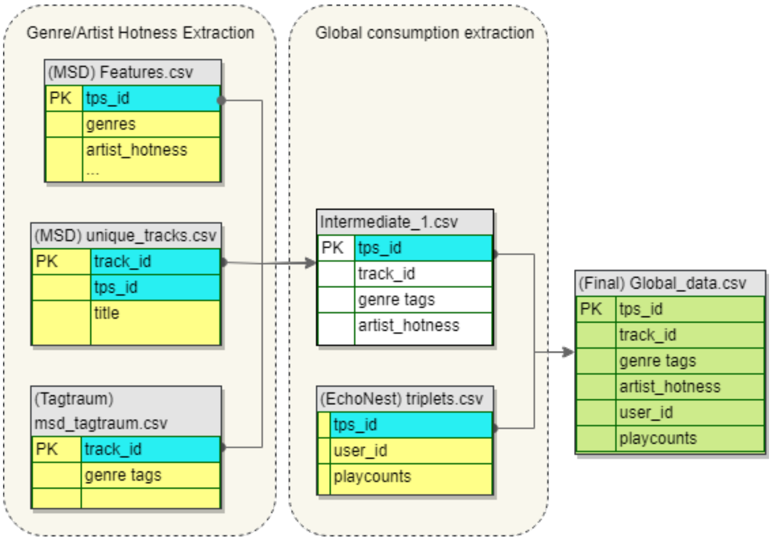


FIGURE 3.3: Data integration for global profile. The datasets are color coded based on their type: yellow represents source datasets, white is the intermediate dataset and green is the final dataset.

All Music genres: The first genre dataset that we used was provided by Hu & Ogihara [41], which classifies MSD songs into ten different genre groups. These genre groups were scraped from All Music Guide website[2]. The number of tracks in each genre is given in Table 3.3. The genre distribution diagram (Figure 3.4) shows a very skewed distribution of genres. The main reason for this is that the classification is broad and two major genres - 'Rock' and 'Pop' - are grouped as a single genre - 'Pop/Rock', thereby including around 70% of the total songs in the same category. Furthermore, each song in this dataset is associated with only one single genre, which is not always the case in the music domain.

TABLE 3.3: Genre distribution of 'All Music Genres' dataset

Genres	Number of songs
Pop/Rock	1,25945
Blues	18,632
Rap	12,469
Country	9,942
Jazz	8,266
R&B	7,101
Electronic	2,243
Reggae	1,392
Unclassified	1,133
International	465
Latin	47

For any system that intends to reflect user profiles, it is necessary that it provides a representation of these profiles, as accurate as possible in order to avoid bias due to misrepresentation. This is indeed dependent on the accuracy and granularity of the genre and artist hotness features for our visualization, and using the abstract genre tags of the All Music dataset might result in a false representation of user profiles. Hence we

[2] http://allmusic.com/, as of March 2018

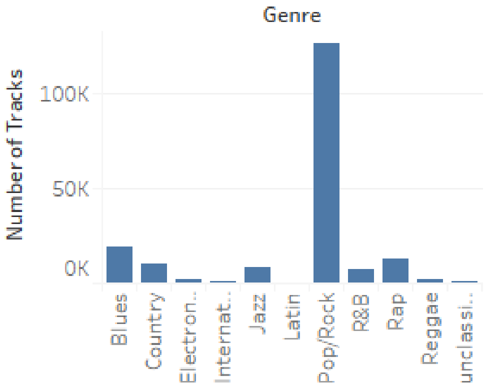


FIGURE 3.4: Distribution of Genres

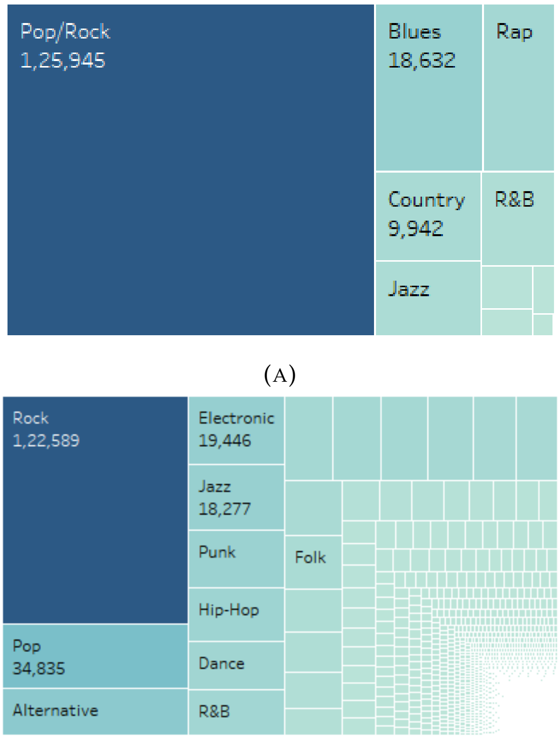
decided to use a second dataset - ’Tagtraum genre annotations’ - provided by Hendrik Schreiber [34].

Tagtraum genres: Unlike All Music Genres, Tagtraum genre annotations are provided for each track of MSD by combining data from multiple sources using majority voting. For details on how the genre information is processed, we refer the reader to their work [34]. The final dataset contained a total of 1000 different genre annotations for 1 Millon tracks. But since these tags were obtained through crowdsourcing, there were a lot of natural language tags that were not exactly genre names and hence the data still required some cleaning. We performed the following preprocessing to the dataset before merging it with MSD:

1. **Standardized genre tags**: Genres with different names referring to same categories were given a standard name. For example, in the original dataset, unclassified tracks were represented as ’Unclassified’, ’Unclassifiable’, ’Other’, ’None’ or ’Uncategorized genre’. We standardized this to a single category name ’Unclassified’.

2. **Standardized item separators**: Genre annotations from Tagtraum were crowdsourced, and hence a uniform format was not used to represent genre combinations. In the existing dataset the following delimiters were used for item separators: ,.+/- . These were replaced by a single delimiter token - ’/’.

3. **Filtered redundancies**: Each track in the dataset is associated with at least two genre tags. However, out of the 1 Million tracks, 1,496 tracks were associated with at least 100 genre tags, and 7,954 tracks were associated with at least 50 genre tags each. On further analysis, we found out that most of these genre tags were redundant or were not actually genres (example: there were tags like "good stuff", "awesome" etc.). We filtered out these genres based on their strength value provided by Tagtraum. A strength value (0 to 1) is assigned for each genre in each track, and it states the relative confidence with which the tracks can be classified in that particular genre. After testing with multiple strength values, a threshold value of 0.2 was chosen, as it provided the least redundancy with the most number of genre annotations. Therefore all genre tags with a strength of 0.2 or more were retained while the rest were eliminated.



(A)

(B)

FIGURE 3.5: (A) Genre distribution in All Music dataset. (B) Genre distribution in tagtraum dataset after pre-processing.

After pre-processing the dataset, we obtained a total of 6,936 different genre tags. As evident from the tree-map (Figure 3.5), the distribution of Tagtraum genres is more fine-grained compared to the genre distributions in All Music dataset.

Finally, after obtaining the refined genre dataset, the last step is to merge the refined Tagtraum genres with the existing MSD dataset. Items in the MSD dataset are indexed based on their song ids whereas Tagtraum dataset is indexed based on track ids. This caused a well-known mismatch error[3], and to fix this issue MSD provided an intermediate dataset *unique_tracks.csv*, that maps between the track IDs and song IDs of MSD. We used this dataset to merge MSD tracks with genre annotations from Tagtraum and All Music genres. Upon merging the three datasets, we found that out of 1 Million tracks of MSD, Tagtraum genres contributed to about 40% of the total tracks. For the rest of the tracks, we used the genre information provided by All Music genres.

3.4.1.2 Artist hotness extraction

The extraction of artist hotness (AH) is straightforward, as this value is provided directly by MSD under the field name *artist_hotttness*. Each song is associated with an AH score which says how 'hot' the artist of the song is. The value is algorithmically computed, and it lies between 0 and 1. We did the following pre-processing for the existing AH value:

[3] https://labrosa.ee.columbia.edu/millionsong/blog/12-1-2-matching-errors-taste-profile-and-msd, as of March 2018

1. **Excluded nulls**: We removed all rows with null AH values. Out of 44,745 artists, 15,992 were retained after the filtering.

2. **Excluded outliers**: As stated in MSD feature description[4], AH typically lies between 0 and 1, with 0 representing the least hot and 1 representing the hottest artist. In the original data set, there were few outliers with values ranging from -28.125 to 702, which we filtered. Thus, out of the remaining 15,992 artists, 32 were removed this way, and a total of 15,960 artists remained.

It is important to note that the rows in MSD are indexed based on songs and not artists. Hence AH values are directly associated with songs based on the artist of the song. Figure 3.6 shows the distribution of artist hotness values for all song ids (*tps_id* in MSD). There were 466 artists with an AH value of 0, and no artist had full hotness (i.e., 1). The highest value of AH is 0.9724 (Kanye West), and the average value is 0.4213.

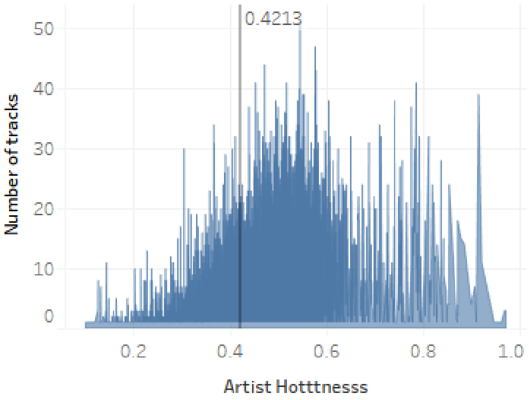


FIGURE 3.6: Distribution of artist hotness (called 'artist_hotttnesss' in MSD) across songs

3.4.1.3 Consumption data extraction

In our visualization, we aim to compare the global consumption pattern with the consumption pattern of the specific user. To obtain such a global profile, we require access to music consumption behavior of as many users as possible, along with their preferred tracks and ratings. Echo Nest provides this data in the name of 'Taste Profile Subset' (TPS) [59], which contains user-song-play count triplets collected from an undisclosed source. TPS is the official user dataset of MSD, and it is the largest music activity dataset made available to researchers. Below are some useful statistics on TPS:

1. The original dataset contains 48,373,586 user-song-play count triplets. Out of these, 1,47,138 unique songs were identified, that matched with MSD songs.

2. Ratings are provided in the form of play counts, and each song in the dataset has been played at least once. Figure 3.7 shows the distribution of play counts across songs. The chart has a long-tail distribution where most of the songs have smaller play counts. Almost 50% of the total songs have 1 to 7 play counts, and 80% of the tracks have 1 to 20 play counts. The average play count for a song is 20.

[4] https://labrosa.ee.columbia.edu/millionsong/pages/field-list, as of March 2018

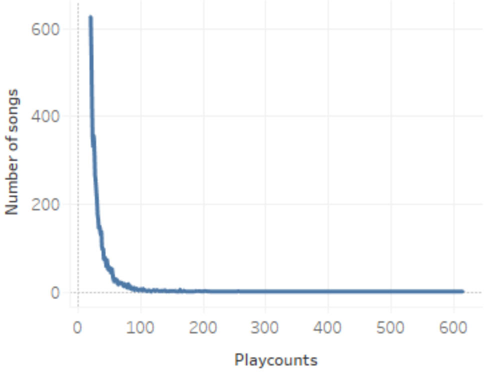


FIGURE 3.7: Distribution of play counts

3. Each user has a minimum of 20 to a maximum of 454 unique song ratings, with an average of 37 ratings per user. The distribution of songs per user is shown in Figure 3.8, and again, the chart has a long tail distribution where fewer users listen to a large number of songs.

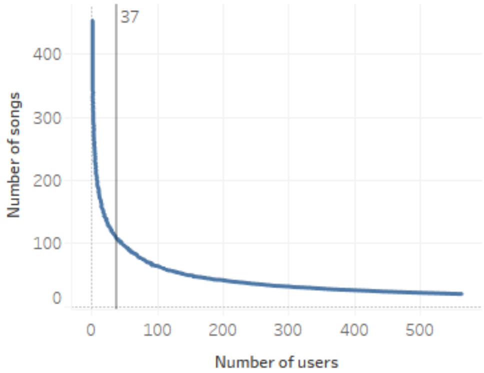


FIGURE 3.8: Distribution of tracks across users

Finally, after merging the TPS with MSD dataset (including the annotated genre and artist hotness information), the final dataset was reduced from 1 million entries to 3,82,589 entries. This reduction was the result of eliminating null entries for artist hotness value, and merging MSD and TPS datasets. The rest of the sections use this final dataset, and hereafter we refer to this dataset as the 'global dataset'.

3.4.2 User Profile Extraction

To build a user profile and to enable comparison with the global profile, we need to obtain a specific user's real-time music listening pattern. Similar to global profile, this entails all the track preferences of the user, and the genre/genre-combinations and artist hotness values of these tracks. Several music APIs are available for researchers that give authorized access to user's listening history. A comparison of these APIs in terms of requirements for our research goal is shown in Table 3.4.

TABLE 3.4: Comparison of existing music APIs for user data extraction.

Music Data APIs	**User consumption data**	**Genre tags**	**Artist hotness (AH)**
MusiXmatch	No	No	No
Last.fm	**Yes**	No	No
iTunes	No	**Yes**	No
Spotify	**Yes**	**Yes**	**Yes**

Among the available APIs only Last.fm and Spotify provides a real-time user consumption data. They both provide a list of user's recently listened tracks in addition to a list of user's top tracks (calculated from the play count of the track).

When it comes to **genre** information, even though last.fm does not directly provide genre tags, it can be obtained indirectly by mapping last.fm tracks with MSD tracks. But this method of genre extraction does not work for tracks that were released after 2011 since MSD data is scrapped only until 2011. This might be an issue during the online evaluation of the system where participants tend to have more recent songs in their profile. Spotify, on the other hand, does not associate each track with genre tags, but it does provide genre tags for individual albums and artists. These genre tags are assigned from a comprehensive list of 1750 genres. A detailed list of all the available genres and sample tracks from these genres can be found in the 'Spotify genre map' [5] provided by Echo Nest. For **artist hotness** value, none of the APIs except Spotify provide this information. In Spotify, artist hotness is referred to as 'Artist Popularity', and it lies between 0 and 100.

Thus, considering the above discussion, Spotify is the only API that is capable of providing the required features. Besides, Spotify is one of the largest music service providers, and hence it is relatively easy to find real users for evaluation. Therefore, we proceed with Spotify for obtaining user consumption data, and in the rest of the sections, we discuss the specific API end-points that were used to extract each of these features.

3.4.2.1 Genre extraction

Spotify does not provide genre information directly for tracks, but instead, it associates genre tags with albums and artists of the track. Out of these, genre tags associated with albums are very sparse, and most of the times the API returns a null array for genres. To minimize null genres, we perform the following three passes for genre extraction for each track:

1. **Album meta-data**: In the first pass, we look into the track's album metadata to get the *album's* genre information. We use Spotify's 'get album' API endpoint [6] which returns the album's genres as an array of strings. If this array is null, it means that the album's genre tag is unknown, and so we proceed to the next pass.

[5] http://everynoise.com/engenremap.html, as of June 2018
[6] https://api.spotify.com/v1/albums/{id}, retrieved May 2018

2. **Artist meta-data**: In the second pass we look for the *artist's* genre tag. For each track of the user that has a null genre, we first obtain the artist id of the track. Using this ID, we extract the artist's metadata using Spotify's 'get artist' endpoint[7]. From this metadata, the artist's list of genres can be obtained as an array of strings. In general, Spotify has more genre tags for artists than albums and hence most of the time this value is available. But in cases where this array is empty, we proceed to the third and final pass.

3. **MSD genres**: In the final pass we look up the genre information in the global dataset that we constructed using MSD in Section 3.4.1. Extraction of genre for a Spotify track from MSD dataset is not straightforward since the track IDs of both these datasets are different. To deal with this issue, we used a mapping dataset provided by 'Acoustic brainz' project[8], which maps MSD IDs to IDs from other music services including Spotify. Based on this mapping, we then extracted the genre for the specific Spotify track by matching it with the corresponding MSD track.

After the above three passes, if a track still doesn't have any associated genre tags, we eliminate this track from our study.

3.4.2.2 Artist Hotness extraction

Spotify associates each artist with a popularity value, ranging from 0 to 100, with 100 being the most popular. This field is the closest equivalent to artist hotness tag of MSD, and it can be extracted from the artist meta-data[9].

Contrary to MSD, artist popularity in Spotify is not calculated directly for each artist but derived mathematically based on the popularity of the artist's individual tracks. The popularity of individual tracks, in turn, is algorithmically derived based on the total number of plays the track has and how recent the plays are. Accordingly, a track/artist with a large number of recent plays will have more popularity than a track/artist with a large number of past plays. In order to enable comparison with global data, we normalize this popularity value to lie in the range 0 to 1.

3.4.2.3 Consumption data extraction

For user consumption data, Spotify API provides two endpoints to choose from - the first one gives a list of user's *recently played tracks*, and the second one gives a user's *top tracks*.

1. **Recently played tracks**[10]: This endpoint returns a user's most recent 50 tracks. A track is included if it is played for more than 30 seconds. One main drawback of this endpoint is that it does not provide the play counts of these tracks. Hence, if a user played the same track five times, it will be shown as five individual tracks.

[7] https://api.spotify.com/v1/artists/{id}, retrieved May 2018

[8] https://acousticbrainz.org/, retrieved March 2018

[9] https://api.spotify.com/v1/artists/{id}, retrieved May 2018

[10] https://api.spotify.com/v1/me/player/recently-played, retrieved May 2018

Since the access is limited to just 50 tracks, this kind of repetition might narrow the perspective on user's preferences.

2. **Top tracks**[11]: This endpoint provides the top 50 tracks for each user of the system. Top tracks are selected based on a calculated *Affinity* value. This value is a measure of the expected preference a user has for a particular track, and it is computed based on user behavior. Since Spotify does not explicitly provide play counts for tracks, using this list gives a better representation of user's profiles. Furthermore, this list does not contain duplicates and hence we can get a broader view of the user's preferences. Hence, we use this endpoint to fetch user consumption data.

Since a user's behavior is likely to shift over time, Spotify provides the above preference data over three time spans: short-term (approximately last 4 weeks), medium-term (approximately last 6 weeks) and long-term (calculated from the beginning of user's profile creation). Choosing the short-term preference data would not be representative of a user's actual taste if there is any bias in the user's recent consumption. Similarly choosing the long term preference data would not provide a good representation of the user's current blind-spots. Hence to achieve a proper trade-off between an accurate representation of a user's preferences, and their blind-spots, we chose medium-term representation.

Finally, for each user for each of the top tracks, we extract the genres/genre-combinations and artist hotness values of the tracks as mentioned in Section 3.4.2.1 and 3.4.2.2. We do this for all 50 top tracks to obtain the final 'user dataset'.

3.5 Frequent genre pattern extraction

After obtaining the global and user dataset, in the next step, we identify the most frequent genres/genre-combinations for each of these profiles. We achieve this by applying **Frequent item-set Mining** algorithm, individually to both global and local dataset. This algorithm identifies the most frequent (i.e., most recurring) genre/genre-combinations, based on the given consumption pattern. Such an identification would enable us to visualize user's blind-spots by highlighting how their most frequent genre consumptions differ from that of the global population.

Several algorithms have been proposed in the literature to implement frequent item-set mining. Some of the most widely used algorithms include a-priori [4], FP-Growth [35], ECLAT [88] and RElim [18]. As a result of its simplicity and efficiency, we use RElim for our research. In this section, we briefly explain the working of RElim and its advantages.

3.5.1 Recursive Elimination Algorithm (RElim)

Recursive Elimination Algorithm (RElim) was first proposed by Christian Borgelt [18] in order to find frequent item-sets. RElim is a simple algorithm that works on *recursive elimination scheme*, where each item is recursively eliminated from the list of items if it is

[11] https://api.spotify.com/v1/me/top/{type}, retrieved May 2018

not individually frequent, i.e., if it appears fewer times than a user-specified minimum threshold (*support* value).

RElim follows a step-by-step elimination of items from the dataset. The steps involved in the pre-processing stage of RElim is shown in Figure 3.9 and discussed below:

- **Step 1**: The transaction database is taken in its original form
- **Step 2**: The frequencies of individual items are determined based on this input.
- **Step 3**: Infrequent items are eliminated based on their user-specified support value.
- **Step 4**: The remaining items are sorted lexicographically in decreasing order of their frequency.
- **Step 5**: The data structure on which RElim operates is built by setting up a list in which each element is grouped based on its leading item and consists of two fields: an occurrence counter and a pointer to a list of its trailing items.

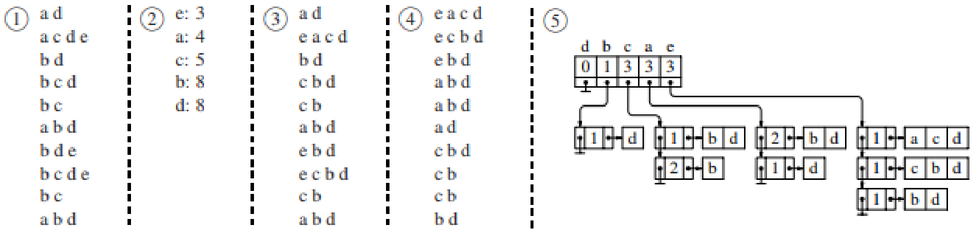


FIGURE 3.9: Pre-processing steps for RElim algorithm [18]

Each item in the data structure is then processed recursively from left to right (Figure 3.10). During each recursion, the elements on the item's trailing list are either discarded or reassigned to the remaining items in the list, and the counters of the corresponding items are updated. This step is repeated for all items in the list, and the final counter values for each item of the list gives the frequency of the item and item-combinations in the dataset. For more details about RElim algorithm and it's pseudo-code we refer the reader to the author's original paper [18] and it's follow-up [19] respectively.

Compared to other existing frequent item-set mining algorithms, RElim is better for the following reasons:

1. Unlike other frequent item-mining algorithms [88, 4, 35] RElim is much easier to understand with simpler steps of pre-processing for input dataset.
2. The performance of RElim is proven to be higher that its alternatives (FP-Growth, ECLAT) with shorter execution time [65, 18].

3.5.2 Implementation

For the Python implementation of RElim, we used the *pymining* package [69], which has a collection of data-mining algorithms implemented in Python. For each global and

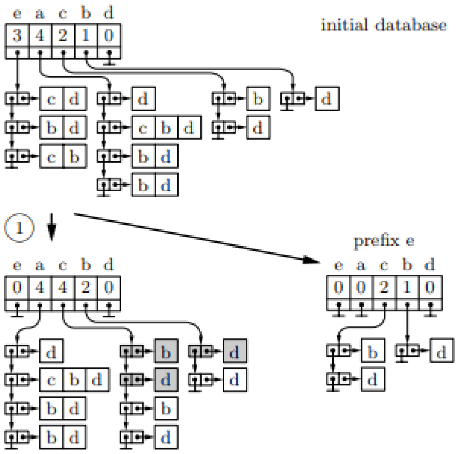


FIGURE 3.10: Procedure of recursive elimination with modification of the transaction lists (left) and construction of transaction lists for the recursion (right) [18]

user data set, this algorithm gives us a list consisting of a set of frequent genres/genre-combinations along with their frequency values (i.e., the number of times the item-set has occurred). For the global dataset, we take the play counts into consideration by replicating each genre a number of times equivalent to the play count of its corresponding tracks.

The minimum support (*min_support* parameter) was set to 2 which means that any genre that occurs less than two times will be eliminated from the list. Since eventually we only retain top-20 most frequent genre/genre-combinations for our visualization, this support value would not make much impact. However, it does reduce the processing time by eliminating insignificant items earlier during pre-processing. Furthermore, since for specific user profile we only gain access to top 50 tracks (using Spotify API), the support value should be small enough so as to still show frequent item-sets for such data. For all these reasons we chose a value of 2 for support.

Table 3.5 shows the top 20 most frequent genre/genre-combinations along with their (normalized) frequencies for the global dataset. Here, the top three globally most frequent genres (Rock, Pop, and Alternative) are in-line with the distribution of these genres in the Tagtraum dataset (Figure 3.5b). Out of these, 'Rock' has the highest preference globally. People also seem to prefer a certain combination of genres more than other individual and combined genres. For example, a combination of Alternative and Rock has a higher frequency compared to Rap or Metal.

Once we have the most frequent genre/genre-combinations, computation of the second dimension for visualization, average artist hotness value is pretty straightforward:

- For each of the top-20 genre/genre-combinations, we first identify their corresponding tracks.
- For each of these tracks, we then obtain their artist hotness values and compute the average.

Table 3.6 shows the average artist hotness values (0 to 1) for the previously obtained top 20 item-sets. Since the data is representative of the global population, we expected that

TABLE 3.5: Top 20 frequent item-sets for global dataset with a support value of 2.

Genres (1-10)	Frequency	Genres (11-20)	Frequency
Rock	0.308	Metal	0.029
Pop	0.108	Rock, Punk	0.029
Alternative	0.075	Rock, Metal	0.028
Alternative, Rock	0.071	Country	0.027
Hip-Hop	0.038	Dance	0.023
Electronic	0.036	Rock, Pop	0.023
Rap	0.032	Alternative, Punk	0.021
Rap, Hip-Hop	0.032	Alternative, Rock, Punk	0.021
R&B	0.032	Latin, Indie	0.020
Punk	0.029	Indie	0.020

the average artist hotness values lie close to the center (0.5) for the global dataset. This is explainable since different users prefer artists from different points in the popularity spectrum. From the table, it is evident that this expected pattern is clearly reflected in the obtained values.

TABLE 3.6: Average artist hotness (AAH) values for the top 20 frequent item-sets for global dataset (support value of 2).

Genres (1-10)	AAH	Genres (11-20)	AAH
Rock	0.56	Metal	0.55
Pop	0.56	Rock, Punk	0.53
Alternative	0.57	Rock, Metal	0.55
Alternative, Rock	0.57	Country	0.55
Hip-Hop	0.56	Dance	0.5
Electronic	0.5	Rock, Pop	0.53
Rap	0.57	Alternative, Punk	0.54
Rap, Hip-Hop	0.57	Alternative, Rock, Punk	0.54
R&B	0.59	Latin, Indie	0.46
Punk	0.53	Indie	0.52

3.6 Choice of Visualization

For our study, we not only aim to understand *if* visualizations are able to convey to users, their consumption pattern, but we also aim to identify *which* visualization is better suited for such a purpose. Hence we *compare* two different visualizations and study their ability to convey to users, their consumption pattern, and blind-spots. The choice of visualizations are based on their ability to satisfy most, if not all, of the following criteria:

1. **To span across multiple dimensions**: The chosen visualization should be capable of representing the frequent genres and genre-combinations, their corresponding frequencies and average AH values.

2. **To span across multiple profiles**: The visualization should be capable of differentiating the global profile from the user's profile.

3. **To represent coverage**: The visualization should be capable of representing the genre distributions clearly, i.e., it should be capable of distinguishing highly frequent genres from the less frequent ones.

4. **To highlight relationships between items**: For genre combinations, the visualization should be capable of representing the interactions between two genres. For example, to represent the genre-combination *'Alternative, Rock'*, the relationship between *'Alternative'* and *'Rock'* genre should be clearly highlighted.

5. **To enable comparison between multiple profiles**: The visualization should enable users to compare their frequent genre/genre-combinations with that of the global profile, and in doing so, it should require minimal cognitive effort from the users.

Based on these criteria, we chose scatterplot as our main visualization and bar-line chart as our base-line visualization. In the following sections, each of these visualizations is discussed in detail.

3.6.1 Visualization 1: Scatterplot

Scatterplot is a type of plot that uses Cartesian coordinates to display values for typically two dimensions of data. The data is represented as a collection of points, with each point having the value of one variable determining its position along the horizontal (x-) axis and the second variable determining its position along the vertical (y-) axis. Traditional scatterplots are capable of representing only two dimensions. However, with the inclusion of visual attributes such as color, size, and shape it is possible to represent up to five dimensions. This serves perfect for our purpose since we require the visualization to represent the following four dimensions:

1. Top 20 genre/genre-combination, ie., the name of the item-set
2. Normalized frequency of the top genre/genre-combination (lies between 0 to 1)
3. Artist hotness value of the top genre/genre-combination (lies between 0 to 1)
4. Profile type (i.e., differentiate 'user' and 'global' profile)

An example of scatterplot visualization used in our study is shown in Figure 3.11a. The attributes associated with each dimension of our visualization is shown in Table 3.7. The most important dimension that we aim to highlight is the frequency of each genre, and we use the size of the bubbles to represent this variable. Therefore, the larger the bubble, the higher the frequency of the genre corresponding to that bubble.

To distinguish between genres we use color hues since it is one of the highly recommended attributes for categorical data [49]. We do not show genre names as labels of bubbles (as it would make the visualization more cluttered), rather we implemented a hover feature, where the genre name, frequency and AH value of a bubble is displayed in a tool-tip when the user hovers over the bubble. We also highlight same genres in both user and global profile on hovering over one of the bubbles corresponding to that genre. This enables users to to compare their profiles with the global profile easily.

Finally, we assign the rest of the dimensions, i.e., AH value and profile type to the horizontal and vertical axes respectively.

From the example in Figure 3.11b, we can infer the following information:

1. For the given user, *Pop* is the most frequently consumed genre since it corresponds to the largest bubble under user ('yours') category of vertical axis.
2. Pop is also highlighted under the global category, which means that it is also globally one of the most (but not *the* most) frequent genre(s).
3. The user's bubbles are generally aligned more towards the right compared to the bubbles in the global category. This means that the user prefers more popular artists compared to the average user of the system.

Finally, it is possible that a genre is in the global profile but not in the user's profile and vice-versa. Genres that are most widely consumed globally but not present in the user's profile are called as the user's *blind-spots*. In the scatterplot, this can be identified by hovering over each bubble in the global category and checking if that genre is highlighted in the user's profile. If the genre is not highlighted in the user's profile, then the genre is not present in the user's profile, and hence it is his/her blind-spot.

TABLE 3.7: Assignment of scatterplot attributes to dimensions

Data dimensions	**Scatterplot attributes**
Genre/Genre-combination Frequency	Size of the bubble
Genre/Genre-combination name	Color of the bubble
Artist hotness value	Horizontal axis
Profile type	Vertical axis

Out of the required criteria mentioned in Section 3.6, our scatterplot visualization satisfies four (1,2,3 and 5) criteria. For representing relationships between items (criteria 4), even though the scatterplot does not highlight relationships in terms of interactions, it does represent the item-set as a single independent element ('*Alternative, Rock*' from the above example). In retrospect, this compromise provides a rationale for the relative ease of use and understandability of the visualization.

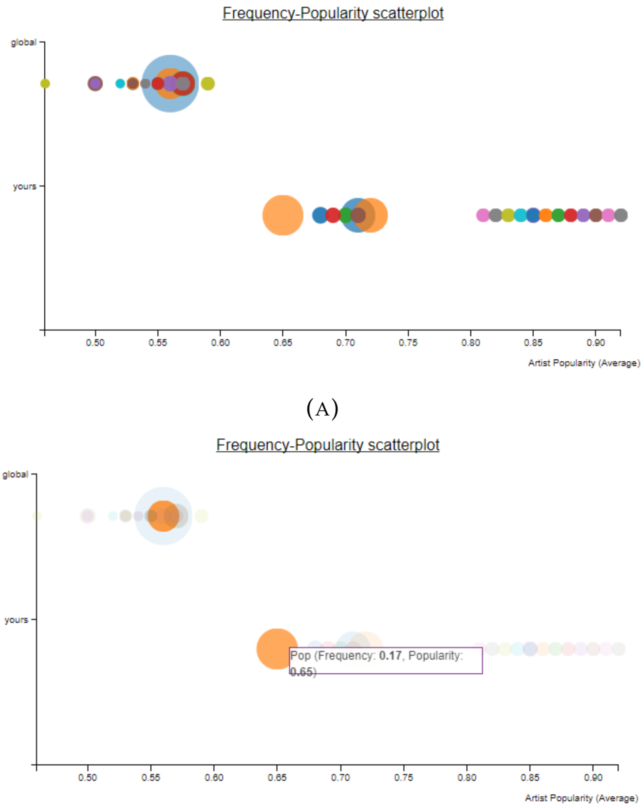
3.6.2 Visualization 2: Bar-line chart (Baseline)

We compare the performance of scatterplot with the base-line visualization bar-line chart. Bar-line chart is a combination of bar chart and line chart, and it can represent up to three variables. A bar chart based visualization was chosen as the baseline for the following reasons:

1. It is proven to be the most compelling and persuasive means (out of a total of 21 means) to convey explanations in recommender systems [37].
2. It is used in existing recommender systems such as Movie Lens[12] to represent user's ratings across genres, and frequency of ratings (Figure 3.12).

A bar chart represents data with rectangular bars, where the height of the bar is proportional to the values they represent. For our purpose, we require that the chart is capable

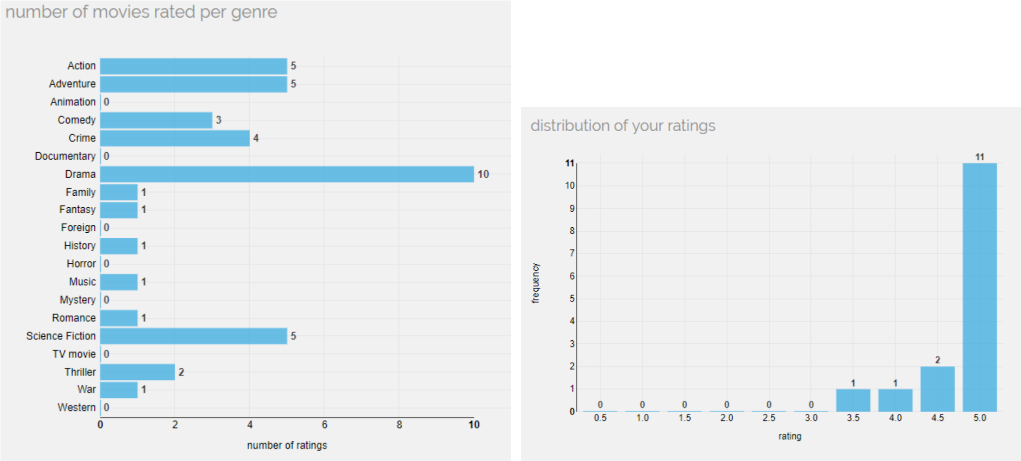
[12] Movie Lens: https://movielens.org/, as of June 2018



(A)

(B)

FIGURE 3.11: (A) An example scatterplot visualization used in the study. The horizontal axis represents average artist hotness value; the vertical axis represents profile type - *yours* (for user's profile) or *global*; the color of the bubble indicates different genres and size of the bubble indicates (normalized) frequency of the genres. (B) On hovering over a bubble, its corresponding genre name gets highlighted along with its frequency and artist hotness value. If the same genre is found in both global and user's profile, it gets highlighted in both the places.



(A)

(B)

FIGURE 3.12: Bar chart being used in existing Movie Lens system (A) to represent the number of movies rated in each genre for a given user (B) to represent the distribution of ratings

of representing all four dimensions - top genre/genre-combinations, the normalized frequency of top genre/genre-combinations, artist hotness value of top genre/genre-combinations and profile type (Section 3.6.1). But with a simple bar chart, one can only represent a maximum of two dimensions - one along the horizontal and the other along the vertical axis. It is, however, possible to include an additional dimension by either including a third (z-) axis or a secondary horizontal or vertical axis. Including a third (z-) axis would give us a pseudo-3-D bar chart, the use of which is frowned upon in research due to its complexity [63]. Hence we extend our chart by adding a secondary axis, more specifically, a secondary *vertical* axis. We use line-chart to represent data points in this axis, since bar-line chart is a standard combination used commonly in Excel [13].

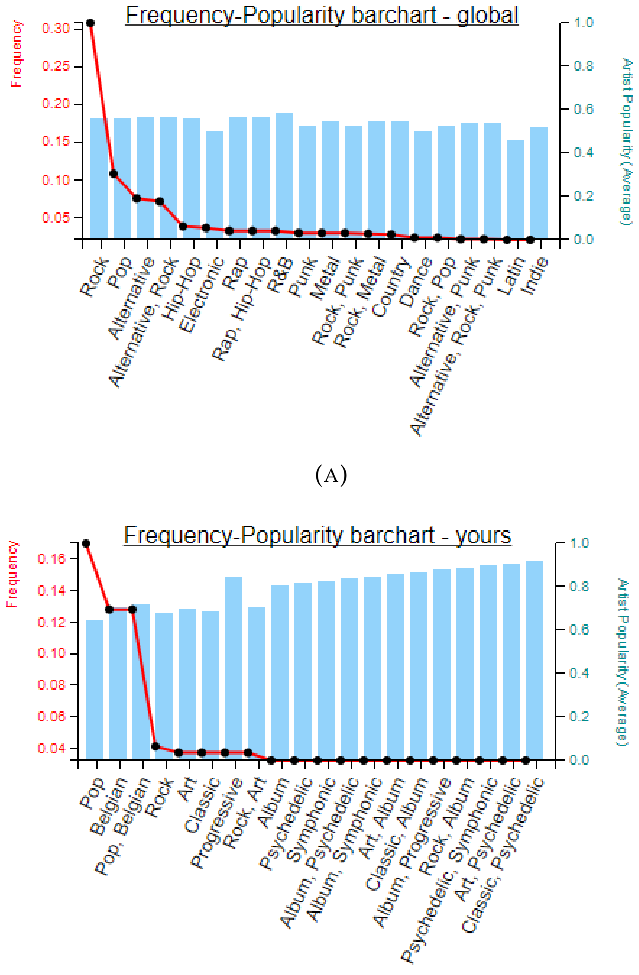
By using two vertical and one horizontal axes, we are able to represent, (a) the genre names along the horizontal axis, (b) the frequency of the genres along the left vertical axis (corresponding to the line-chart), and (c) the average artist hotness values along the right vertical axis (corresponding to bar-chart). Table 3.8 shows the dimensions corresponding to the respective attributes of the chart. Unlike scatterplot, a single bar-line chart does not accommodate both user and global profiles. Hence we use two different bar-line charts to separate the profiles (Figure 3.13a and 3.13b). This separation might make it difficult to enable comparison between both profiles. But on the other hand, it accounts for the simplicity of the bar-line chart in comparison to the scatterplot.

TABLE 3.8: Assignment of bar-line chart attributes to dimensions

Data dimensions	**Bar-line chart attributes**
Frequency of genre/genre-combination	Left vertical axis (lines)
Name of genre/genre-combination	Horizontal axis
Average artist hotness value	Right vertical axis (bars)
Profile type	Chart title

Figure 3.13 shows the equivalent bar chart visualization for a sample user. More specifically, Figure 3.13a shows the global profile and 3.13b shows an example user's profile.

[13] Combining chart types, adding a second axis: https://www.microsoft.com/en-us/microsoft-365/blog/2012/06/21/combining-chart-types-adding-a-second-axis/, as of June 2018



(A)

(B)

FIGURE 3.13: (A) Bar-line chart for global profile (B) Bar-line chart for the user's profile

It is evident from the figure that globally *Rock* is the most preferred genre, while for that specific user it is *Pop*. Furthermore, few genres like *Alternative*, while widely preferred globally, is not present in the user's profile, and hence these are the user's *blind-spots*. When it comes to artist hotness value, the bars are generally longer in user's profile compared to the bars in the global profile. This says that the user tends to prefer artists with higher popularity compared to the average user of the system.

4 Evaluation 1 - to study user's understanding of visualization

4.1 Introduction

In the previous chapter, we described in detail, the steps involved in the extraction and visualization of consumption pattern. With our visualizations, we provide an overview of a user's consumption pattern (individual profile) in comparison with the consumption pattern of other users of the system (global profile). Consequently, we intend to answer our three research questions (Section 1.2) restated below:

1. Are visualizations effective in conveying to users, their consumption pattern and blind-spots? [RQ1]
2. Which visualization is the most effective in representing to users, their consumption blind-spots? [RQ2]
3. Does user's understanding of their profile correlate with their intention to explore their blind-spot genres, and if it does, what is the strength of such a correlation? [RQ3]

To test the extent to which our system answers the above research questions, we perform an online evaluation of the system. Based on the three research questions, we divide the whole evaluation process into two conceptual stages. In this chapter, we discuss stage 1 of evaluation, where we evaluate user's understanding of both bar-line chart and scatterplot visualizations (for RQ1 & RQ2). In the next chapter, we discuss stage 2 of evaluation where we study the correlation between user's music exploration pattern and their understanding of their consumption pattern. It is important to note here that such a classification of the evaluation process is introduced solely for the purpose of better representation of concepts, and from participant's perspective, the whole evaluation is staged as a single experimental session.

In the following sections, we explain the experimental set up and results of stage 1 of evaluation. We first explain the experimental design (Section 4.2), variables (Sections 4.3 & 4.4) and research hypotheses (Section 4.5). This is followed by a brief explanation of the steps involved in the evaluation in Section 4.6.We then brief about the materials in Section 4.7, participants of the study in Section 4.8 and measures in Section 4.7, followed by the results for each of our hypothesis in Section 4.10. Finally, we end this chapter with a detailed discussion of the obtained results in Section 4.11.

4.2 Design

In stage 1, we focus on studying a user's understanding of their consumption pattern and blind-spots with both bar-line chart and scatterplot. We compare both visualizations to see which one provides a better representation of such information. We used a **within-subjects** repeated measures design, where each participant was presented with both scatterplot and bar-line chart. For each of the visualization, the participant was asked to answer a few questions that test their understanding of their consumption pattern and blind-spots. In order to minimize order effects, we performed counterbalancing of the order of appearance of the visualizations.

4.3 Independent Variable

For each user, we show both types of visualizations (bar-line chart and scatterplot), and we study the effectiveness of each of these visualizations in increasing the understandability of a user's consumption pattern and blind-spots. Hence **type of visualization** is our independent variable. An example of both the visualizations are shown in Figure 4.7 for bar-line chart and Figure 4.3 for scatterplot.

4.4 Dependent Variables

We have two dependent variables :

1. **Correctness of understanding**: Understandability of visualization is measured by asking users to answer questions about information represented in the visualization. These questions test a user's understanding of their consumption pattern and blind-spots.
2. **Confidence**: In addition to measuring the user's actual understanding, we also measure the perceived understandability for both the visualizations. These are self-suggested confidence scores provided by the user for each question about their consumption pattern and blind-spots. This value says how confident the users are in their answers about their consumption pattern and blind-spots, and it is measured on a five-point Likert scale from very low confidence to very high confidence.

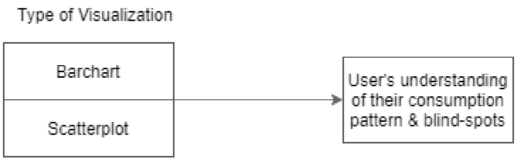

Type of Visualization
Barchart
Scatterplot
User's understanding of their consumption pattern & blind-spots

FIGURE 4.1: Conceptual model for Stage 1

4.5 Hypotheses

We devised a total of five hypotheses to answer our research questions, out of which, four are validated in this chapter. These are:

- **H1**: Users are able to answer questions about their consumption pattern more accurately with scatterplot than with bar-line chart.
- **H2**: Users have more confidence in their answers about their consumption pattern for scatterplot more than bar-line chart.
- **H3**: Users are able to answer questions about their blind-spots more accurately with scatterplot than with bar-line chart.
- **H4**: Users have more confidence in their answers about their blind-spots for scatterplot more than bar-line chart.

4.6 Procedure

Each participant goes through five steps of evaluation. Each of these steps is discussed below:

1. In the first step, participants read a consent form, where a brief overview of the experiment is provided, and they offer their consent to take part in the study.
2. In the second step, we obtain the user's basic demographics and information about their music consumption and music background. Table 4.1 shows the list of questions asked to the users in this stage.

TABLE 4.1: Demographic questionnaire

ID	Questions
D1	What gender do you most identify with?
D2	What is your age group?
D3	Please fill in your profession/domain
D4	How often do you listen to music?
D5	How often do you search for new music?
D6	What is your musical background?

3. In the next step, users are asked to log in with their Spotify account. Once a user logs in with his/her account we collect the user's top 50 tracks using Spotify API's *top tracks* endpoint[1]. We then use frequent pattern mining algorithm on the genres of these tracks to compute the user's top frequent genres, and interaction between these genres (i.e., genre-combinations). To know more about how the algorithm works, we refer the reader to Section 3.5.
4. In the fourth and fifth step, users are presented with either a bar-line chart or a scatterplot representation of their consumption pattern. Each visualization displays the user's top genre/genre-combinations alongside the global profile.

[1] https://developer.spotify.com/documentation/web-api/reference/personalization/get-users-top-artists-and-tracks/ - retrieved May 2018

Users are first required to read a brief set of instructions explaining the visualization. The exact instructions provided for bar-line chart and scatterplot are given below:

Bar-line chart instructions: "*In this page, you will see two bar-line charts. The chart on the left shows the music consumption behavior of ALL users of the system, and the chart on the right shows YOUR music consumption behavior. Using these two charts, you could compare your music preferences with the music preferences of other users of the system. For each chart:*

- *The horizontal axis represents the most listened genres.*
- *The left vertical axis represents the frequency: how many tracks have been listened in that genre. This value lies between 0 to 1 (higher = more listens). In the chart, this is represented as black dots on a red line.*
- *The right vertical axis represents the 'average artist popularity': how popular an artist of a song is. Again, the values are between 0 to 1 (higher = more popular)."*

Scatterplot instructions: "*This diagram shows you a scatterplot representation of the type of music YOU listen to - 'your' category of the vertical axis, compared to the type of music other users of the system listen to - 'global' category of the vertical axis. Here are a few things you need to know:*

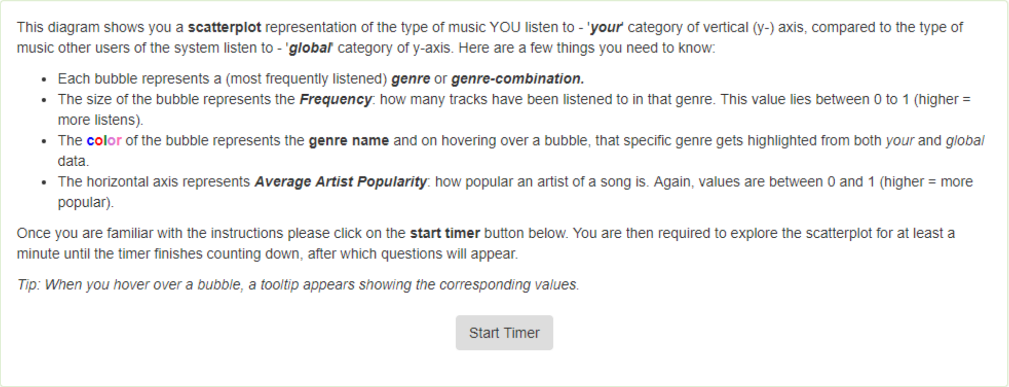
- *Each bubble represents a (most frequently listened) genre or genre-combination.*
- *The size of the bubble represents the frequency: how many tracks have been listened in that genre. This value lies between 0 to 1 (higher = more listens).*
- *The color of the bubble represents the genre name, and on hovering over a bubble, that specific genre gets highlighted for both your and global data.*
- *The horizontal axis represents the 'average artist popularity': how popular an artist of a song is. Again, the values are between 0 to 1 (higher = more popular)."*

Once the user gets familiar with the instructions, he/she is asked to click on the 'start timer' button present at the end of the instructions. This starts a countdown timer for one minute, during which time, the user is asked to examine the visualization (Figure 4.2). This ensures that all users spent a minimum amount of time trying to read the visualization.

Once the timer counts down to zero, a set of questions appear for the user to answer. The questions are designed to be answered directly by looking at the visualization. The correctness of user's answer to these questions is a proxy for the user's understanding of the visualization. A brief description of the choice of questions is given below:

Choice of questions: The questionnaire is designed in such a way that, for each visualization, they evaluate user's understanding of the system in all four aspects - global consumption pattern, user's consumption pattern, user's blind-spots and artist hotness values. More particularly, we ask users to identify the top first and second genres for each of the four aspects. In other words, we ask users to identify:

(a) globally first and second most consumed genres



This diagram shows you a **scatterplot** representation of the type of music YOU listen to - '***your***' category of vertical (y-) axis, compared to the type of music other users of the system listen to - '***global***' category of y-axis. Here are a few things you need to know:

- Each bubble represents a (most frequently listened) ***genre*** or ***genre-combination.***
- The size of the bubble represents the ***Frequency***: how many tracks have been listened to in that genre. This value lies between 0 to 1 (higher = more listens).
- The **color** of the bubble represents the **genre name** and on hovering over a bubble, that specific genre gets highlighted from both *your* and *global* data.
- The horizontal axis represents ***Average Artist Popularity***: how popular an artist of a song is. Again, values are between 0 and 1 (higher = more popular).

Once you are familiar with the instructions please click on the **start timer** button below. You are then required to explore the scatterplot for at least a minute until the timer finishes counting down, after which questions will appear.

Tip: When you hover over a bubble, a tooltip appears showing the corresponding values.

FIGURE 4.2: Once the 'start timer' button is clicked, a count-down timer starts for one minute, during which time, users are asked to explore the visualization. After the counter stops, the questionnaires appear.

(b) the user's first and second most consumed genres

(c) the user's first and second highest blind-spot genres

(d) artist hotness values of these genres (i.e., the user's most consumed and blind-spot genres)

Furthermore, for each question, we ask users to provide their confidence in their answer, which gives the user's perceived understanding of the visualization. Finally, for each of the visualization, we also ask users to provide their feedback about the visualization. More specifically, we ask users "what they liked/disliked about the visualization".

The above process is repeated for both visualizations, and in order to minimize the effect of familiarity with profile and learning effects, we perform a strategic split of questions and ask half of the questions for the first chart and the other half for the second chart. In doing so, we also counterbalance the order of these questions, as to which half of the questions goes to which chart. By performing such a split, we were also able to considerably decrease the survey time, thereby reducing user's fatigue and boredom. Specific details on how we split the questions can be found in Appendix A.

4.7 Materials

We use both bar-line chart and scatterplot as the means to convey user's consumption behavior. These visualizations were created using D3.js javascript visualization library [25]. An example of both the charts for a sample user is shown in Figure 4.3 and Figure 4.7. For a single user, the same consumption data is represented in both scatterplot and bar-line chart, however, due to dimensionality limitations of the latter we use two bar-line charts to represent the same information that can be represented by a single scatterplot (Section 3.6.2).

Between two different users, the content displayed in the chart varies depending upon their consumption pattern. More specifically, between two different users, data shown in the *user's* region of the chart varies (i.e., the region marked 2 in Figure 4.3 for scatterplot and the whole of Figure 4.4b for bar-line chart). However, the global consumption data remains the same for all users (the region marked 1 in Figure 4.3 for scatterplot, and the whole of Figure 4.4a for bar-line chart).

The survey was developed using Python Flask framework [30].

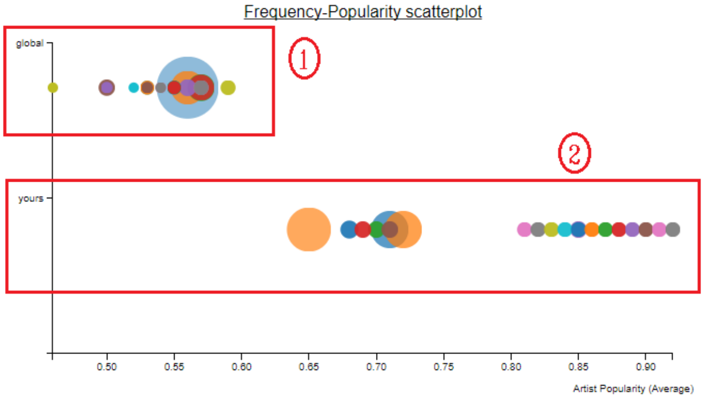
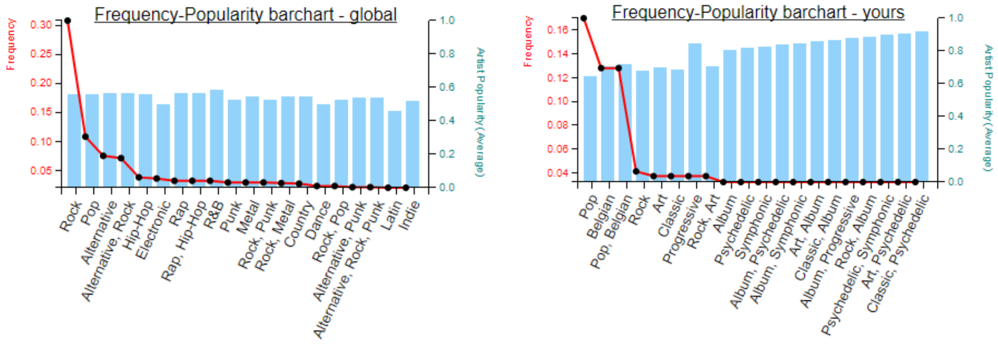


FIGURE 4.3: Example scatterplot used in the study: section 1 shows the global consumption pattern and 2 shows the user's pattern



(A)

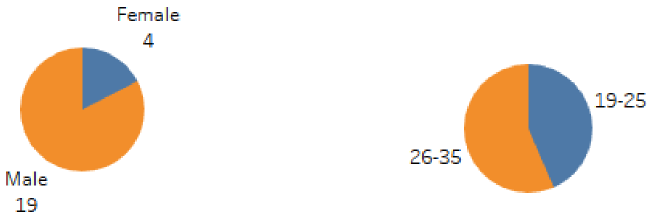
(B)

FIGURE 4.4: Example bar-line chart used in the study: (A) represents global consumption pattern and (B) represents user's pattern

4.8 Participants

There were a total of 23 participants recruited from a European Technical University. 83% of the participants (n = 19) were male and 17% female (n = 4) (Figure 4.5a). Out of the 23 participants, 19 had a computer science background, with 17 MSc students and 2 Ph.D. students (Table 4.2). Hence, there is a high probability that they are already exposed to visualizations in one or the other forms. 43% of the participants (n = 10) were of the age-group 19-25 and 57% of the participants (n = 13) were between 26-35 years of age (Figure 4.5b).

Participants had a diverse music consumption behavior and musical background. Out of the 23 participants, 48% (n = 11) stated that they listen to music between 2 to 4 hours a



(A)

(B)

FIGURE 4.5: Participant demographics by (A) gender, and (B) age group

TABLE 4.2: Participant demographics by profession

Profession	Number of participants	Percentage
MSc computer science	17	73.9%
PhD computer science	2	8.6%
MSc chemical engineering	1	4.3%
MSc life sciences	1	4.3%
MSc bio-mechanical	1	4.3%
Automation engineer	1	4.3%

day and 39% (n = 9) stated that they listen to music for more than 4 hours a day (Figure 4.6). When it comes to musical background, most of the participants had advanced (43%), Basic (26%) or no expertise (26%) in music. Tables 4.3 and 4.4 show detailed demographics about participant's musical background and their frequency in seeking new music respectively.

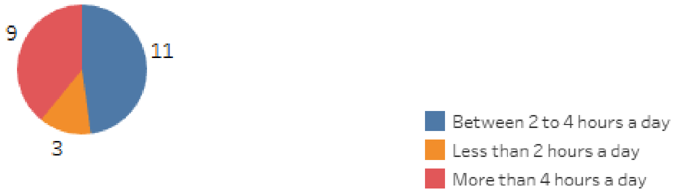


FIGURE 4.6: Participant demographics by their music listening frequency

4.9 Measures

In our study, we quantify the user's understanding of visualizations based on the answers they provides for questions about their consumption pattern and blind-spots. More specifically, for each question that the user answers, we assign a score based on the correctness of their answer. Therefore, for each question:

- If the answer is right (i.e., if the user identifies the right genre or provides the right artist hotness value), a score of 1 is assigned
- If the answer is partially right (i.e., if the user identifies the second best answer), a score of 0.5 is assigned. For example, for a question that asks to identify the highest blind-spot, if the user answers with his *second* highest blind-spot, he gets a score of 0.5.

TABLE 4.3: Participant demographics by musical background

Musical background	Number of participants	Percentage
Professional - professional musician (conductor, composer high level instrument player), music conservatory student, audio engineer, etc.	1	4.3%
Advanced - regular choir singing/ amateur instrument playing/ remixing or editing music with computer etc.	10	43.4%
Basic - lessons at school/ reading music magazines, blogs, etc.	6	26.08%
None - no particular interest in music related topics	6	26.08%

TABLE 4.4: Responses to the question, "How often do you search for new music?"

Music search frequency	Number of participants	Percentage
Always	2	8.6%
Very often	5	21.7%
Often	6	26.08%
Sometimes	4	17.4%
Rarely	6	26.08%
Never	0	0%

- All other answers are treated as wrong, and the user gets a score of 0.

This level of scoring provides a fine-grained analysis of results compared to the standard binary scoring (0-1) and yet manages to keep the analysis simple. Besides, this scoring is in-line with the scoring used in a previous study [62] and therefore enables comparison of results.

Once we have obtained all our scores, we used Mann-Whitney U-test to examine the statistical significance of our results.

4.10 Results

In this section, we describe the results of our user-centric evaluation aimed at measuring user's understanding of the visualizations. We posit our results in accordance with the four hypotheses as stated in Section 4.5, and for each hypothesis, we state if the results provide enough evidence to accept the hypothesis.

Understandability 1: Consumption pattern

H1: Users are able to answer questions about their consumption pattern more accurately with scatterplot than with bar-line chart

For each participant, we compute the scores for their answers, about their first and second highest consumed genres and the artist hotness values of these genres. Table 4.5 shows the mean scores of all participants, for identifying their first and second most consumed genres for both bar-line chart and scatterplot. The mean values are relatively higher for bar-line chart than for scatterplot for both first and second place. However, these differences are not statistically significant as shown below:

TABLE 4.5: Mean (std) scores of all participants for identifying their first and second most consumed genre. Scores lie between 0 to 1 with 1 being the highest score representing maximum understanding.

	Bar-line	**Scatter**
1st most consumed	1.00 (0)	0.95(0.14)
2nd most consumed	1.00 (0)	0.92(0.29)

- First most consumed genre: Mann-Whitney U-test, U-value = 60 (critical value = 33), at p<0.05 - **Not significant**.
- Second most consumed genre: Mann-Whitney U-test, U-value = 60.5 (critical value = 33), at p<0.05 - **Not significant**.

Table 4.6 shows the mean scores for identification of artist hotness values of the user's first and second most consumed genres for both visualizations. The mean value is higher for scatterplot for the artist hotness of first most consumed genre, and for bar-line chart for the second most consumed genre. However, the results are, again, not statistically significant:

1. AH of first most consumed genre: Mann-Whitney U-test, U-value = 55 (critical value = 33), at p<0.05 - **Not significant**.
2. Second most consumed genre: Mann-Whitney U-test, U-value = 60.5 (critical value = 33), at p<0.05 - **Not significant**.

TABLE 4.6: Mean (std) scores of all participants for identifying the artist hotness values of their first and second most consumed genre. Scores lie between 0 to 1 with 1 being the highest score representing maximum understanding.

	Bar-line	**Scatter**
AH of 1st most consumed	0.87 (0.31)	1.00 (0)
AH of 2nd most consumed	1.00 (0)	0.92 (0.29)

Thus, **we have no support for H1**.

Confidence 1: Consumption pattern

H2: Users have more confidence in their answers about their consumption pattern for scatterplot more than bar-line chart.

Confidence values are directly obtained from users for each of their answers about their consumption pattern and blind-spots. User's average confidence in their answers about their first and second most consumed genres are shown for both charts in Table 4.7.

TABLE 4.7: Mean (std) confidence scores of all participants for identifying their first and second most consumed genre. Scores lie between 1 to 5 with 5 representing maximum confidence.

	Bar-line	**Scatter**
1st most consumed	4.60 (0.49)	4.00 (0.89)
2nd most consumed	4.18 (0.75)	4.45 (0.96)

The values suggest that, for the identification of the first most consumed genre, users have higher confidence in bar-line chart than scatterplot. For the second highest genre, they have a higher confidence with scatterplot than bar-line chart. These trends, however, are not significant:

- (Confidence) First most consumed genre: Mann-Whitney U-test, U-value = 36 (critical value = 33), at p<0.05 - **Not significant**
- (Confidence) Second most consumed genre: Mann-Whitney U-test, U-value = 59 (critical value = 33), at p<0.05 - **Not significant**

We also looked at the confidence scores provided by users for identifying the artist hotness values of their first and second most consumed genres (Table 4.8). The scores seem to adhere to the previous pattern (Table 4.7), i.e., users have higher confidence for bar-line chart for identifying the artist hotness of first most consumed genre, and higher confidence for scatterplot for the second most consumed genre. The results of statistical significance tests are shown below:

- (Confidence) Artist hotness of first most consumed genre: Mann-Whitney U-test, U-value = 29 (critical value = 33), at p<0.05 - **Significant**
- (Confidence) Artist hotness of second most consumed genre: Mann-Whitney U-test, U-value = 50 (critical value = 33), at p<0.05 - **Not significant**

Thus for identification of the artist hotness value, for the first most consumed genre, users report *higher* confidence for bar-line chart than for scatterplot. This result **contradicts our hypothesis (H2)**.

Understandability 2: Blind-spots

H3: Users are able to answer questions about their blind-spots more accurately with scatterplot than with bar-line chart

TABLE 4.8: Mean (std) confidence scores of all participants for identifying the artist hotness of their first and second most consumed genre. Scores lie between 1 to 5 with 5 representing maximum confidence. * represents significant results.

	Bar-line	**Scatter**
1st most consumed*	4.90 (0.28)	4.27 (0.64)
2nd most consumed	4.30 (0.67)	4.70 (0.49)

To compare a user's understanding of their blind-spots for both the visualizations, we compute the scores for their answers about their first and second highest blind-spots. Table 4.9 shows the mean scores for identification of the first and second highest blind-spots for both visualizations.

TABLE 4.9: Mean (std) scores of all participants for identifying their first and second highest blind-spot genres. Scores lie between 0 to 1 with 1 representing maximum understanding.

	Bar-line	**Scatter**
1st blind-spot	0.66 (0.49)	0.68 (0.46)
2nd blind-spot	0.82 (0.34)	0.83 (0.39)

From the above table, it is evident that users have a higher average score with scatterplot than with bar-line chart. This seems to suggest that users have a higher understanding with scatterplot for identifying their blind-spots than with bar-line chart. The obtained results are not statistically significant:

- First highest blind-spot: Mann-Whitney U-test, U-value = 66 (critical value = 33), at p<0.05 - **Not significant**
- Second highest blind-spot: Mann-Whitney U-test, U-value = 61 (critical value = 33), at p<0.05 - **Not significant**

A similar trend is observed in the mean scores for the artist hotness values of first and second highest blind-spots. Table 4.10 shows the average understandability scores for the artist hotness of first and second highest blind-spots for both charts. We see again that the average values are slightly higher for scatterplot compared to bar-line chart. The results of Mann Whitney U-test are shown below:

- Artist hotness of first highest blind-spot: Mann-Whitney U-test, U-value = 66 (critical value = 33), at p<0.05 - **Not significant**
- Artist hotness of second highest blind-spot: Mann-Whitney U-test, U-value = 61 (critical value = 33), at p<0.05 - **Not significant**

The statistical significance test fails for this observation too, thereby deeming the results insignificant at p<0.05. Hence, we **found no support for H3**.

TABLE 4.10: Mean (std) scores of all participants for identifying the artist hotness of their first and second highest blind-spot genres. Scores lie between 0 to 1 with 1 representing maximum understanding.

	Bar-line	**Scatter**
1st blind-spot	0.66 (0.49)	0.70 (0.46)
2nd blind-spot	0.81 (0.34)	0.83 (0.39)

Confidence 2: Blind-spots

H4: Users have more confidence in their answers about their blind-spots for scatterplot more than bar-line chart.

We computed the average confidence values for all participants for their answers about their blind-spots. The results are shown in Table 4.11 for identifying the top two blind-spot genres and Table 4.12 for identifying the artist hotness values of these genres for both the charts.

TABLE 4.11: Mean (std) confidence scores of all participants for identifying their first and second highest blind-spots. Scores lie between 1 to 5 with 5 representing maximum confidence.

	Bar-line	**Scatter**
1st blind-spot	3.80 (0.93)	4.20 (0.90)
2nd blind-spot	4.30 (0.82)	3.90 (0.67)

Results from Table 4.11 suggest that users tend to show higher confidence for their blind-spot genre identification with scatterplot for the first genre and bar-line chart for the second genre. Statistical significance of the results are shown below:

- (Confidence) First highest blind-spot genre: Mann-Whitney U-test, U-value = 46.50 (critical value = 33), at p<0.05 - **Not significant**
- (Confidence) Second highest blind-spot genre: Mann-Whitney U-test, U-value = 43.50 (critical value = 33), at p<0.05 - **Not significant**

The confidence scores for identification of artist hotness values of the top two blind-spot genres (Table 4.12) seem to correlate with the above results. This time both the results are statistically significant:

- (Confidence) artist hotness of first highest blind-spot genre: Mann-Whitney U-test, U-value = 33 (critical value = 33), at p<0.05 - **Significant**
- (Confidence) artist hotness of second highest blind-spot genre: Mann-Whitney U-test, U-value = 16.5 (critical value = 24), at p<0.01 - **Significant**

This suggests that, for identification of their first highest blind-spot, users have higher confidence with scatterplot and for identification of their second highest blind-spots users have higher confidence with bar-line chart. This result only partially satisfies our hypothesis that scatterplot performs better than bar-line chart for identification of blind-spots and hence we **reject H4**.

TABLE 4.12: Mean (std) confidence scores of all participants for identifying the artist hotness of their first and second highest blind-spots. Scores lie between 1 to 5 with 5 representing maximum confidence. * represents significant results.

	Bar-line	**Scatter**
1st blind-spot*	4.25 (0.62)	4.80 (0.40)
2nd blind-spot*	4.90 (0.30)	4.00 (0.50)

4.11 Discussion

In this section, we discuss implications of our results in the light of our first two research questions, and delineate the post hoc analyses of results.

RQ1: Are interactive visualizations effective in conveying to users, their consumption pattern and blind-spots?

Keeping in mind the main aim of our visualizations - which is to increase user's understanding of their profile - we first looked into the scores that the users obtained for questions about their consumption pattern and blind-spots. The average score for all their answers was **8.7** out of 10 which seem to suggest that users gained a good understanding of their profile from the visualizations.

However, in a user-centric study such as ours, where we examine the effectiveness of visualizations, it is quite possible that user's *actual* understanding of the system contradicts their *perceived* impression of the system (i.e., what they actually know vs. what they think they know about the system). Hence to provide a fair evaluation it is important that we consider both these aspects in our study, and verify if the results complement each other. Keeping this in mind, we further analyzed user's perceived understanding of the system based on their self-suggested confidence scores. The average confidence score for all users for all their answers about their consumption pattern and blind-spots was **4.37** (standard deviation = 0.39). This value is well above the threshold of 3.5 for agreement on a 5-point Likert scale [80, 39], which shows that visualizations gave users a good perceived understanding of the profiles.

Therefore, our results do suggest that users have a good perceived and actual understanding of their profile from visualizations. Therefore, visualizations are, indeed, a good medium to convey users consumption pattern and blind-spots.

RQ2: Which visualization is the most effective in representing to users, their consumption patterns and blind-spots?

To answer RQ2, we assessed each of the visualizations with respect to their ability to convey the user's consumption pattern and blind-spots. For each visualization, we aggregated the scores that the users obtained for correctly identifying their consumption pattern and blind-spots. Based on these scores, and the user's self-suggested confidence values, we compared the performance of visualizations in conveying user's consumption pattern and blind-spots. The results show a specific trend in the user's understanding of their profile, as given below:

1. **Understandability - Consumption pattern**: The average scores that the participants obtained for questions about their consumption pattern were *higher for bar-line chart than scatterplot* (H1 in Section 4.10). But the difference in means was not statistically significant, and we were unable to confirm our results.

2. **Confidence - consumption pattern**: The average confidence scores provided by the participants for questions about their consumption pattern were inconsistent for their first and second positions. That is, to identify their *first most consumed genre*, users are *more confident with bar-line chart* (**significant**), and to identify their *second most consumed genre*, they had *higher confidence with scatterplot* (**not significant**) (H2 in Section 4.10). Here, the results are significant for the first position. Hence we were able to confirm that, at least for the identification of their first most consumed genre, users prefer bar-line chart to scatterplot.

3. **Understandability - Blind-spots**: The average scores that the participants obtained for questions about their consumption pattern were *higher for scatterplot than bar-line chart* (H3 in Section 4.10). But the scores were not statistically significant, and hence we were unable to confirm our results.

4. **Confidence - Blind-spots**: Again, the average confidence scores provided by the participants, for questions about their blind-spots, were inconsistent for their first and second positions. But this time, with a reverse trend. That is, to identify their *first highest blind-spot*, users are *more confident with scatterplot* (**significant**), and to identify their *second highest blind-spot*, they are *more confident with bar-line chart* (**significant**) (H4 in Section 4.10). Both the scores were significant and thus, the results are inconclusive as to which visualization is most effective for the identification of blind-spots.

Our results seemingly suggest a pattern in the performance of visualization. That is, to identify their consumption pattern, users prefer bar-line chart, and to identify their blind-spots they prefer scatterplot. To confirm this pattern, we need more concrete evidence. Therefore, we performed a post-hoc analysis with the comments that users provided for each of their visualizations. The following section discusses this analysis and explains how it supports our results.

Post-hoc analysis

To get further insights into the observed results we analyzed the comments that users provided during the assessment of each visualization.

We first extracted the significant comments from the pool of comments by performing a manual parsing. Specifically, we retained comments that resonated well with a maximum number of users (i.e., comments that were frequently made), and removed polarizing comments. For example, out of the 23 participants, 6 participants agreed that scatterplot enabled better comparison between global and user's data, and 3 participants supported this by agreeing that comparison was difficult in bar-line chart. Therefore, we retained this comment. Tables 4.13 and 4.14 show the top comments extracted this way, along with the number of users who agreed with the comment - for bar-line chart and scatterplot respectively.

TABLE 4.13: User's feedback for bar-line chart along with the number of users who explicitly agreed with the comment.

Bar-line chart	
Pros	**Cons**
Simple, well-quantified, clean interface (5) Easy to read (3)	Confusion between two vertical axes (5) Comparison between global and user's data was difficult (3)

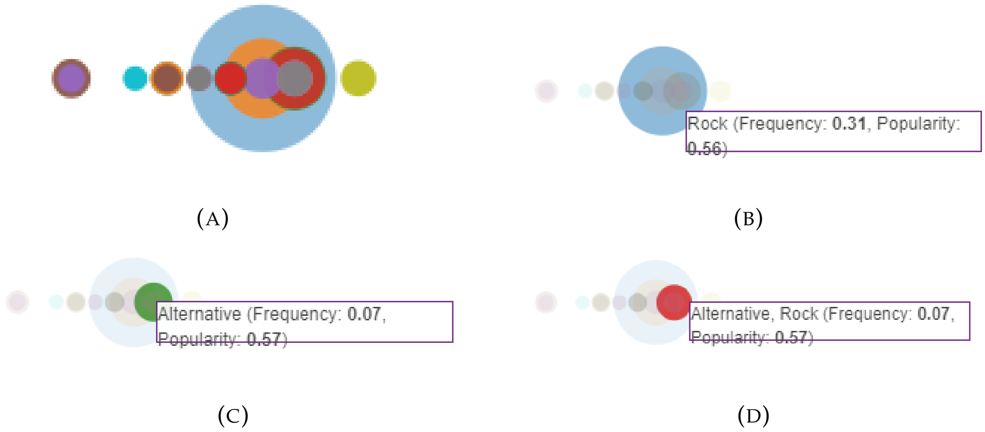
TABLE 4.14: User's feedback for scatterplot along with the number of users who explicitly agreed with the comment

Scatterplot	
Pros	**Cons**
Easy comparison between global and user's data (6) Intuitive (4)	Overlapping bubbles were hard to read (6) Multiple colors were confusing (4)

In general, users agreed that bar-line chart was easier to read and that it was well-quantified, while scatterplot was easier for comparison between the user and global data. This perception seems to explain why users scored higher with bar-line chart for questions about their consumption pattern, and higher with scatterplot for questions about their blind-spots - which requires them to compare global and user data.

One main drawback of scatterplot, as mentioned by several users, is that the overlapping of bubbles made the chart harder to read. Keeping this in mind, and further scrutinizing user's data, we found that *Rock* being the highest blind-spot for most of the users (15 out of 23 participants) is also the biggest and most prominent bubble in the global data with fewer overlaps (Figure 4.7b). On the other hand, for the second highest blind-spot, the genres (*Alternative* for 11 users and *Alternative, Rock* for 8 users) are closely cluttered and overlapped onto each other (Figures 4.7c & 4.7d). Such an overlap might be a reason why users had higher confidence with scatterplot for identification of their first highest blind-spot, but lower confidence for their second highest blind-spot.

Therefore, although users did, in fact, identify their blind-spots correctly with scatterplot, such an overlap of bubbles tends to lower their confidence, in their answers about their blind-spots.



(A)

(B)

(C)

(D)

FIGURE 4.7: Scatterplot representation of global consumption pattern: (A) represents all global data (B) highlighted 'Rock' genre, which is the highest blind-spot for 15 out of 23 users (C) & (D) highlighted 'Alternative' and 'Alternative, Rock' genres, which is the second highest blind-spot for 11 out of 23 and 8 out of 23 users.

Therefore, from the post-hoc analysis, we were able to provide reasoning to our observed results, and further strengthen the possibility that, **users prefer conventional bar-line charts for the identification of their frequent genres**. **For the identification of their blind-spots, users prefer scatterplot**, since it enables better comparison between global and user's data. However, these observed results are not conclusive from the experiment.

4.12 Limitations

In this stage of evaluation, we faced specific scenarios that impose restrictions on the validity of the study. We delineate these limitations in this section:

Construct validity: In the evaluation of hypotheses 1 to 4, we consider the correctness of user's answers to questions (about their consumption pattern and blind-spots) and their confidence scores for the questions, as a proxy to measure the actual and perceived understandability of visualizations. This method is in-line with an existing research [62] that measures the performance of visualizations and hence using the same metric accounts for a better comparison for our study. However, there could be other ways to measure the understandability of the visualization which could be addressed in future research.

Internal validity: Internal validity of a study is increased when the effects of confounding variables (more than one possible independent variables) are addressed. In our study, while measuring the understandability of both bar-line chart and scatterplot visualizations, it is possible that the interactivity of both the visualizations had an effect on the understandability of visualizations. Furthermore, user-specific aspects such as a user's prior knowledge of visualizations, perception of colors, etc., might have also contributed to the high understanding scores of visualizations.

External Validity: Firstly, when it comes to the data used in our experiment, the global consumption data obtained from Million Song Dataset's taste profile subset (TPS), is available only until the year 2011. The EchoNest API that provided access to this information was shut down (on May 31, 2016) [2] before the experiment was conducted, and since then there is no known alternative source to extract this information. Hence it is quite possible that recent changes in consumption trends are not reflected in our global profile.

Secondly, to answer RQ1, which asks if visualizations are better at conveying to users their consumption pattern and blind-spots, we only study the performance of two types of visualizations (Bar-line chart and scatterplot). The choice of these two visualizations was done carefully based on the number of dimensions they could represent (Section 3.6). Accordingly, we eliminated a few common visualizations (such as chord diagram, heat-maps, radar chart, pie-chart and bubble charts) during analysis, citing their inability to represent all the required five dimensions (user type, genre name, genre frequency, average artist hotness value, interaction between the genres). It is possible that there are other visualizations that increase or decrease user's understanding of their profile. Future works could focus on other means of representing this data.

[2] The Echo Nest - Wikipedia Entry - as of March 2018

5

Evaluation 2 - to study user's music exploration

5.1 Introduction

In the first stage, we compare two visualizations by quantifying a user's understanding of their profile for each of the visualization. We use this quantified information in stage 2 to further analyze the impact that such an *understanding of their profile* has on the user's *intention to explore/seek diverse content*. However, in doing so, we abstract the *means* to obtain this understanding, and rather deem understanding as an independent entity. Such a setup, we believe, would help us to better represent the main aim of our system, which is by any means, to expose users to hidden content thereby "priming" them towards diverse exploration (nudge theory [57]).

In the following sections, we explain the experimental design (Section 5.2), dependent (Section 5.4) and independent variables (Section 5.3), research hypothesis (Section 5.5), procedures (Section 5.6), materials (Section 5.7), results (Section 5.10) and discussion (Section 5.11) for stage 2 of evaluation.

5.2 Design

We perform a **correlation study** to analyze the relationship between a user's understanding of their profile, and their music exploration pattern.

5.3 Independent variable

For all participants, we measure if their understanding in their consumption pattern and blind-spots has an impact on the user's exploration of their blind-spot genres. Hence a user's **correctness of understanding** of their profile is the independent variable. This value is directly computed for each user from their scores in stage 1 of evaluation (Section 4.4).

5.4 Dependent Variable

Exploration of blind-spot genres: Each user's exploration of their blind-spot genres is quantified as an exploration factor (EF_{bs}). This value is obtained based on the proportion of genres and tracks the user has explored in their blind-spot category. Blind-spot category is the category of genres that are most preferred globally but not by the user.

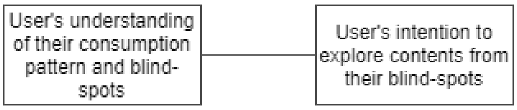

User's understanding of their consumption pattern and blind-spots
User's intention to explore contents from their blind-spots

FIGURE 5.1: Conceptual model for Stage 2

5.5 Hypothesis

H5: Users who have a higher actual understanding of their profile (i.e., their consumption pattern and blind-spots), explore their blind-spot genres more.

5.6 Procedure

After users evaluate both visualizations in Chapter 4, they enter the exploration phase. Here the users perform the following two steps:

1. In the first step, users are provided with a list of genres from their top frequent and blind-spot categories (1 of Figure 5.2). From these genres, users are first asked to select any genre/genre-combinations that they might be interested in listening to at the moment. Based on these genres, a list of songs are recommended (using Spotify's recommender API[1]) in the panel below (2 of Figure 5.2). From this list, users are then asked to listen to the songs that they find interesting. They could also "add" a song to their favorites list if they like the song and want to save it for later (3 of Figure 5.2). Once they think they have explored enough of one genre, they can go back and repeat the process for different genre/genre-combinations. In order to get a more complete impression of their exploration pattern, users repeat the above process for at least five genre/genre-combinations before they could proceed further.

   The main aim of this phase is to study the user's exploration pattern. For each user, we store, (a) the genre/genre-combinations they choose, (b) the tracks they play, (c) the tracks they add from those genres and (d) the amount of time they spent in each genre.

2. Once users have explored different genres, in the final step, users fill in a post-stage assessment questionnaire, that measures the efficiency of visualizations and interfaces. All questions are provided in the form of statements for which users provide their agreement using a five point likert scale. Table 5.1 provides the list

[1] https://developer.spotify.com/documentation/web-api/reference/browse/get-recommendations/ - retrieved May 2018

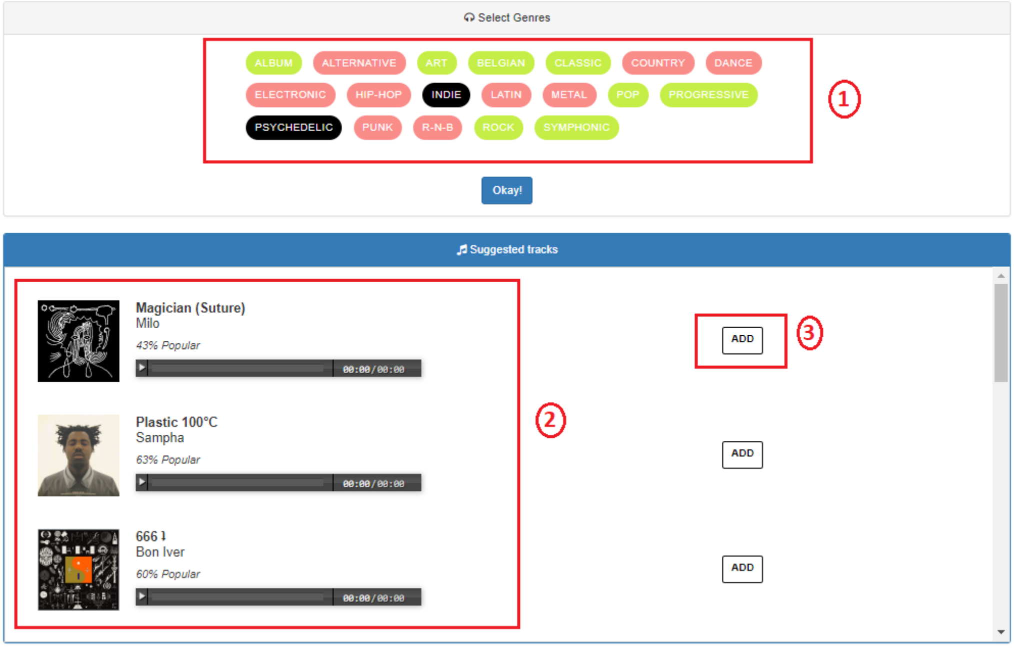


FIGURE 5.2: Exploration phase

of all questions asked in this stage. The questions are chosen in such a way that they capture user's experience, especially with respect to four main aspects of visualizations - *perceived ease of understanding*, *perceived ease of interaction*, *perceived usefulness* and *perceived interest* in visualizations [39]. Precisely, through S5 & S6 we test the perceived ease of understanding, through S8 & S9, perceived ease of interaction, through S10, perceived usefulness and through S11, their perceived interest in visualizing their profile. These aspects provide further insights about user's perception on the effectiveness of our visualizations.

TABLE 5.1: Post-stage assessment questionnaire

ID	Questions
S1	'Artist hotness' value influenced the type of music I listened to
S2	Genre color influenced the type of music I listened to
S3	The visualizations influenced the type of music I listened to
S4	The visualizations made me more confused about the type of music I want to listen to
S5	I became familiar with the visualizations quickly
S6	It requires a lot of effort to understand the visualizations
S7	The visualizations provided the right amount of information
S8	Interacting with the visualization was useful
S9	Interacting with the visualization was clumsy
S10	I see value in using visualizations to show my consumption pattern
S11	I would like to see similar visualizations in Spotify, Netflix, Youtube, etc.

5.7 Materials

Our exploration interface was inspired from Spotify's old genre-mixing interface[2] as shown in Figure 5.3. We adapted this interface, and added color codes to differentiate between frequent and blind-spot genres. Green represents a user's frequent genres and red represents their blind-spot genres (Figure 5.4).

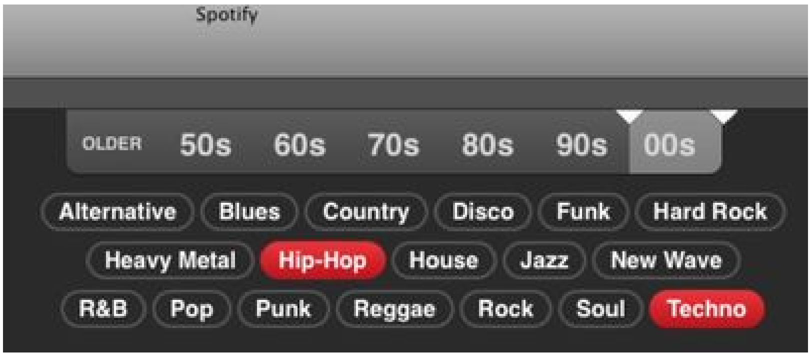


FIGURE 5.3: Original interface from Spotify that lets users to select multiple genres

For recommendation of songs, we use Spotify's recommendation API [3]. This API takes as input the seed genres (chosen by the users), and recommends top songs from these genres.

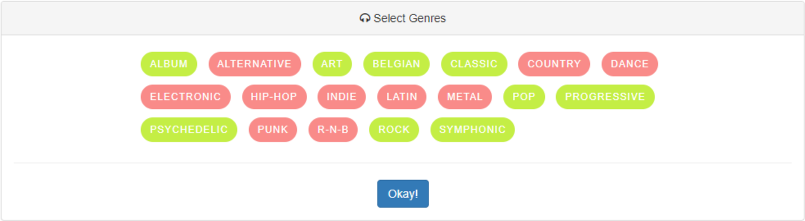


FIGURE 5.4: Exploration interface used in the study

5.8 Participants

Both stage 1 and stage 2 of evaluation are staged as a single experimental session, and therefore the same 23 users who participated in stage 1 continued to stage 2 of evaluation. For detailed demographics we refer reader to Section 4.8.

[2] https://community.spotify.com/t5/Live-Ideas/Radio-Bring-back-selecting-multiple-Genres-amp-Time/idi-p/1013- retrieved May 2018

[3] https://developer.spotify.com/documentation/web-api/reference/browse/get-recommendations/ - retrieved May 2018

5.9 Measures

In the exploration phase, we quantify a user's exploration of their blind-spot category by computing an *exploration factor* (EF_{bs}) for each user. The exploration factor is given as:

$EF_{bs} = N_{bs} * w_{bs}$,

where, N_{bs} is the (normalized) *number of genres* listened by the user from their blind-spot category, weighted by w_{bs}, which is the *proportion of songs* listened by the user from their chosen genres in their blind-spot category. This weight is indeed assigned by dividing the total number of songs listened by the user in their blind-spot category, by the total number of songs listened by the user in all the categories (i.e., their frequent, blind-spot and bridge categories, where frequent category contains the genres that users highly consume, and bridge category is when a user combines genres from their frequent and blind-spot category).

To compute the correlation between the exploration factor and user's understanding we used Spearman's correlation, since it is non-parametric and we do not make assumption of normality in our data distribution.

5.10 Results

In this section, we describe the results for our final hypothesis (hypothesis 5) corresponding to stage 2 of evaluation.

Exploration

H5: Users who have a higher understanding of their profile (i.e., their consumption pattern and blind-spots), explore their blind-spot genres more.

We correlated a user's exploration factor in their blind-spot category (EF_{bs}) with the user's actual understanding of their profile, given by the total scores they obtained for all the questions about their consumption pattern and blind-spots. We found a significant medium Spearman's correlation of **0.44** (significant at $p<0.05$) between the user's actual understanding of their profile and their exploration in the blind-spot category. The correlation has a **moderate** effect size according to Cohen's conventions[44]. This result confirms our hypothesis that users who have higher understanding of their profile, explore their blind-spot genres more, and therefore **we accept H5**.

5.11 Discussion

In this section, we discuss the implications of the results obtained in stage 2 of evaluation, and explain how our study can be extended to answer our final research question.

RQ3: Does user's understanding of their profile correlate with their intention to explore their blind-spot genres? What is the strength of the correlation?

We performed a correlation analysis between the user's understanding of their profile and their intention to explore the unexplored regions of their profile (i.e., their blind-spot category). Our results showed a significant positive correlation (0.44, Spearman's correlation at $p<0.05$).

To further verify the significance of this association, we confirmed that this correlation is exclusive to the blind-spot category, and not observed in frequent and bridge categories. For this, we first computed the exploration factor for the other two categories (frequent - E_f, and bridge category - E_b) in the same way as the exploration factor for blind-spot category. These factors are given as:

- $EF_f = N_f * w_f$, for frequent category
- $EF_b = N_b * w_b$, for bridge category,

where N_f and N_b are the (normalized) number of genres listened by the user from frequent and bridge categories respectively. The weights w_f and w_b are the proportion of songs listened by the user in their frequent and bridge categories. For each category, this is computed by dividing the number of songs listened in that category by the total number of songs listened in all the categories.

For each of the two categories, we then computed the correlation between the obtained exploration factor and the user's understanding of their profile (obtained from their total scores for all their answers). Table 5.2 shows the correlations between the exploration factor of the other two categories (frequent - E_f, and bridge category - E_b) and user's understanding of their profile. The results show that user's exploration in frequent category has a negative correlation with the understanding of their profile (significant at $p<0.05$), and their exploration in bridge category has a very small positive correlation (not significant). This result reinforces our previous conclusion by confirming that the positive correlation between user's understanding of their profile and their exploration in blind-spot category is exclusive.

TABLE 5.2: Spearman's correlation between exploration factor in frequent and bridge genre categories, and user's actual understanding of their profile.

	EF_f	EF_b
Actual understanding	-0.45	0.01

Finally, while comparing user's understanding and their exploration, we are well aware that we only consider their **actual** understanding of their profile. This is because, we believe that, during the exploration phase where users tend to act subconsciously, their behavior is highly influenced by what they actually know rather than what they think they know (i.e., their *perceived* understanding). To confirm our assumption, we verified if there is a stronger correlation between user's exploration factor and their *perceived* understanding of their profile - obtained from their confidence scores. Table 5.3 shows the correlation between user's exploration factor in all three genre categories and their perceived understanding of their profile. We find no significant correlation between the two factors thus proving that user's *actual* understanding of their profile has a much higher influence on their exploration pattern compared to their *perceived* understanding.

TABLE 5.3: Spearman's correlation between exploration factor (in blind-spot, frequent and bridge genre categories), and user's perceived understanding of their profile.

	EF_{bs}	EF_f	EF_b
perceived understanding	0.07	0.08	-0.1

Thus, from the above discussion, we conclude that **user's understanding of their profile, indeed, correlates with their exploration in their blind-spot genres.**

5.12 Additional Observations

In addition to addressing our main research question, our experiment paved way for some interesting observations that are worth mentioning here.

User Preferences

With our exploration interface, we provided control to users to combine genres from their frequent and blind-spot categories for exploration. Given such a control, we did not expect an abrupt evolution in user's genre preferences but rather a gradual shift in preferences from their frequent to blind-spot genres. In other words, we expected that users would prefer to explore genre-combinations from their bridge category more than genres purely from blind-spot category.

Table 5.4 shows the total number of *new* genre/genre-combinations listened by the users in all three possible category combinations. It is evident from the table that users explore a large number of genres from bridge category compared to the other two combinations. This observation is crucial, especially for a diversity-aware recommender system that aims to break the filter-bubble, since it implies that users tend to seek diverse items when it has some elements of their preferences. In other words, users prefer a gradual exposure to content outside their filter-bubble than a sudden and surprising exposure.

Furthermore, from the results of RQ3, we already observed that user's exploration in bridge category is not correlated with their understanding of their profile. Considering this fact, we could further infer that, irrespective of their understanding in their profile, users tend to show interest in exploring their bridge category. Usage of different

TABLE 5.4: Total number of new genre/genre-combinations explored by the user.

Blind-spot genres	**Blind-spot combinations**	**Bridge genres**
38	14	46

color codes to represent the two genre categories in the exploration interface (green for frequent genres and red for blind-spot genres), might in part, have prompted such a behavior.

User control

From literature (Section 2), we concluded that one main issue of providing user control is that, when unaware of the motive, such an autonomy could backfire, and users might effectively shrink their filter-bubble. In our system, we address this issue by providing a degree of awareness to users about their filter-bubble, before giving them the control. As evident from user's exploration pattern, such a set up has significantly steered users towards diverse consumption. Furthermore, if applied rightly, using our exploration interface a user could either expand their filter bubble (by exploring their blind-spot genres) or stay in their bubble (by just exploring their frequent genres), but they can, by no means, exacerbate their bubble.

Currently, none of the existing music recommender systems provide such a control to users where they could combine multiple genres. Spotify did provide a part of this feature, where they allowed users to combine multiple genres, but users were not able to differentiate between their frequent and blind-spot categories. Even though the feature is now withdrawn for unknown reasons, it is evident from support forums[4] that users are still interested in restoring the feature.

5.13 Limitations

In this stage of evaluation, we faced certain scenarios that impose restrictions on the internal and external validity of the study. We delineate these limitations in this section:

Construct validity: In the evaluation of hypotheses 5, we measure user's exploration in their blind-spot genres based on the *number of genres* they explore and the *number of tracks* they explore in each genre. It is possible that considering other aspects of exploration such as the number of minutes each track has been listened, the number of tracks users actually liked, etc., could provide more insight into user's exploration pattern.

Internal validity: When it comes to studying visualizations and exploration, there is a possibility that user's exploration in blind-spot genres is influenced more by the color

[4] https://community.spotify.com/t5/Live-Ideas/Radio-Bring-back-selecting-multiple-Genres-amp-Time/idi-p/1013 - as of May 2018

codes and layout of the genre categories than their actual understanding. Studying the effect of visualizations on user's exploration might provide concrete support for our results.

External validity: When studying the correlation between a user's exploration and their understanding, we compute the exploration factor (E_f) based on the user's genre exploration during the exploration phase. By allowing users to explore music immediately after exposing them to both the visualizations, we were able to track their immediate intention to explore diverse music. Our results do not suggest that such an intention (to explore diverse content) is perpetual. However, by showing a moderate positive correlation between user's exploration and their understanding in their profile, our study provides a good starting point for further analysis in order to verify if such an intention of users to explore diverse music is long-lasting or transient, and if it is dependent on other factors such as the user's visual memory of the displayed visualizations, mood, etc.

6

Conclusion

In this thesis, we devised a method to study the effectiveness of visualizations in increasing user's understanding of their unexplored regions (i.e., their *consumption blind-spots*), and in turn, increasing their exploration in these unexplored regions. We tested our method with Spotify users, and obtained positive results in terms of the effectiveness of visualizations in (a) increasing user's understanding of their profile and (b) nudging them to explore diverse content. In this section, we return to our three main research questions posed in Chapter 1, and address them individually in the light of our proposed approach and conducted evaluations.

6.1 RQ1: Are visualizations effective in conveying to users, their consumption pattern and blind-spots?

In Chapter 3, we set out to expose users to their consumption pattern and blind-spots using two interactive visualizations (bar-line chart and scatterplot). RQ1 addresses the general question as to whether, using visualizations, it is possible to effectively convey information about a user's consumption pattern and blind-spots. To address this research question, in Chapter 4 we performed an online evaluation to test the effectiveness of both our visualizations in increasing user's understanding of their profile (i.e., their consumption pattern and blind-spots). Our results showed that, on average, users scored higher for answering questions about their profile using both the visualizations (score of 8.7 out of 10). Hence, we concluded, in Section 4.11, that our interactive visualizations were indeed effective in increasing user's both *actual* and *perceived* understanding of their profile.

In addition to assessing the *understandability* of a visualization, throughout research, several other perceived aspects, such as ease of interaction, perceived usefulness, etc. have been considered as a measure of quality of visualizations. During our post-stage assessment survey (Chapter 5), we measured three of these aspects: *perceived ease of interaction*, *perceived usefulness* and *perceived interest in visualizations*. Table 6.1 shows the mean responses obtained for each of the three aspects. A value above the threshold value of 3.5 represents agreement on a 5-point Likert scale [80, 39], and our results show that on average, the responses lie above 3.5 for all the three aspects (with 99.99% confidence, 3.891 SE_x, SE_x= standard error of the mean). This result further supports the effectiveness of our visualizations, and based on these results, we address our RQ1:

we have demonstrated that visualizations are effective in conveying to users, their consumption pattern and blind-spots.

TABLE 6.1: Mean (std) values of user ratings and their confidence values for different perceived aspects of visualizations. The values lie between 1 to 5, with 5 being the highest rating representing strong agreement.

Aspects	Mean (std)
Perceived ease of interaction	4.00 (0.76)
Perceived interest	4.40 (0.50)
Perceived usefulness	4.40 (0.70)

6.2 RQ2: Which visualization is the most effective in representing to users, their consumption patterns and blind-spots?

To answer RQ2, in Chapter 4, we performed a within-subjects study to test the user's understanding of both bar-line chart and scatterplot visualizations. For each visualization, users answered a set of questions about their consumption pattern and blind-spots. Based on the correctness of their answers, and their self-suggested confidence scores, we set out to select the visualization that best represents a user's profile. But from the results of our primary analysis, we soon realized that there is no single best visualization, but both visualizations have certain key aspects that enabled them to perform better in different scenarios. For example, we observed that the conventional bar-line chart helped users to better identify their consumption pattern. But for the identification of blind-spots, conventional visualizations seemed ineffective compared to the more complex scatterplot visualization. From our post hoc analysis of user comments, we were able to reason that this complexity of scatterplot, in part, was ruled out by its ability to enable comparisons between the user and global profiles, thereby making it a more comfortable choice for identifying blind-spots.

Therefore, from the results of our primary and post hoc analysis, we address RQ2 with some certainty: Bar-line chart is more effective in conveying information about user's consumption pattern, and scatterplot is effective in conveying blind-spot information.

6.3 RQ3: Does user's understanding of their profile correlate with their intention to explore their blind-spots?

For our third research question, we studied the participant's exploration pattern in Chapter 6, by first exposing them to their most frequent and blind-spot genres, and then studying their music exploration pattern. For each participant, we computed an exploration factor based on the number of genres and the number of tracks listened by him/her in their blind-spot categories.

Our results showed that user's exploration in their blind-spot genres (EF_{bs}) had a **moderate** positive correlation (Spearman's correlation, r = 0.44 at $p<0.05$) with their understanding of their profile. The results were further supported by the fact that this correlation is exclusive to the blind-spot category, and not observed in the exploration of user's frequently consumed genres or their bridge genres (i.e., the combination of blind-spot and frequent genres). Thus by combining these results, we are able to address RQ3 with confidence: users with a higher understanding of their consumption pattern and blind-spots explore their blind-spot genres more.

6.4 Limitations

In this thesis, we faced certain circumstances that imposed restrictions on our approaches, thereby limiting the generalizability of our experiment. In this section, we delineate these limitations and delimitations, with the intention that any future attempt at extending or replicating the study is aware of the scope of our results.

External validity: In our study, all our participants (n=23) were students of a technical university with prior exposure to visualizations. While this provided for a more accessible sample, it might have reduced the ability to generalize our results to a wider population which is unlikely to be so heavily made up of students.

When it comes to the perception of colors used in our visualizations and interfaces, none of our participants openly reported any difficulties in color perception (such as color blindness) during the online evaluation. However, according to statistics[1] 8% of men and 0.005% of women around the world suffer from color-blindness. Since our visualizations (especially scatterplot) relies on colors for identification of genres, it might prove difficult for people with color perception anomalies to get a good understanding of their profile. Hence when extending our work to a real-world music system (like Spotify), alternatives such as using mono-chromatic color coding should be considered.

Finally, it is important to note that our visualizations were built specifically to represent the chosen *music* features (genre, artist hotness etc). Therefore, in order to adapt our visualizations to other domains, such as movies, news, etc. research still needs to be done on choosing the right features corresponding to the domain. It is also possible that based on these features, other visualizations would be more suitable for different domains. Thus, even though the basic idea of visualizing user's consumption pattern and comparing it with the global consumption pattern is applicable to all domains, the exact path to achieve this might vary depending upon the feature set of the individual domain.

[1] http://www.colourblindawareness.org/colour-blindness/

6.5 Future work

Our approach is, to the best of our knowledge, the first approach that provides blind-spot aware content exploration to users. The experiments show positive results in terms of the effectiveness of interactive visualizations in increasing user's diverse exploration. Future works could overcome some of the limitations of the system, and eventually extend our system to incorporate diversity-aware personalization. Below we discuss some of the possible directions.

User profiling & recommendation: With our system, we are able to quantify a user's awareness and exploration in their blind-spot genres. Future research could focus on incorporating this information to create unique user profiles based on these quantified values and eventually enable recommender systems to provide learned diversification of recommendations.

One main challenge that needs to be overcome here is to differentiate between a user's lack of awareness and their lack of interest in the content that remains unexplored. But since our existing system quantifies a user's understanding (awareness) of their blind-spots, we would effectively be able to highlight the specific blind-spot genres that the user is interested in, by studying his/her exploration in those genres. As a user's awareness and exploration changes, the system could continuously update the user's profile.

Interactive visualization: Another area that research would focus on is to provide more control to users while exploring the visualizations. Currently, we only visualize two dimensions (frequency and artist hotness values). However, by identifying and incorporating other significant music features, users might get a better understanding of their profile. Furthermore, depending on the user's musical background and expertise, different users might prefer to see in their visualizations, different aspects of their music consumption pattern. Future research could take this into account to provide control to users to enable them to select their preferred attributes for visualization. Providing user controlled visualization has already been implemented in a recent work [46], where the authors use visualizations to display different attributes of recommended songs. This could be incorporated in our system to provide control to users so that they could filter and visualize their most preferred attributes of their frequent and blind-spot genres.

Contextual factors: Finally, in our current system we do not consider the effects of individual traits on a user's exploration of genres. This drawback could be overcome by incorporating contextual features, such as mood, time of the day, etc., while examining user's exploration pattern. In the long run, this would enable us to create a more accurate representation of the evolution of user's tastes.

6.6 Concluding remarks

The objective of this research has been to help users to break their filter-bubble, by making them aware of their choices, and providing them with an opportunity to explore new content. Although several approaches to break filter-bubble have been suggested, most of the works either fail to keep users in the loop during this process, thereby restricting user's autonomy, or provide control and autonomy to users, but with the likelihood of exacerbating their bubbles.

In our thesis, we presented a two-step approach to deal with this issue, where, in the first step, we make users aware of their filter-bubble by visualizing their consumption pattern alongside the consumption pattern of other users of the system. After increasing user's awareness of their profile, in the second step, we provide control to users to explore new items. Our experiments showed that our visualizations were effective in increasing user's understanding of their profile, thereby indirectly nudging users to explore diverse content. These findings suggest that it is possible to break a user's filter-bubble by increasing the user's awareness of their choices, and providing control to explore new item-sets.

The effectiveness of our visualizations in increasing user's understanding about their profile, provide a first step towards the adoption of interactive visualizations in the domain of recommender systems to convey hidden information about user profiles. The multitudinous features available in different domains (such as music, movies, news, etc.) of recommender systems lends itself particularly well to the effective use of interactive visualizations, giving users an opportunity to control and explore new information, and recommender systems an opportunity to learn from user behavior. Given our results, we hope that future research can focus more on approaches to break filter-bubbles without having to compromise user's autonomy.

A

Counter-balancing the order of visualizations and questionnaires

In the within-subjects study, we compare the performance of two visualizations (bar-line chart and scatterplot) to identify the visualization that best conveys the user's consumption pattern and blind-spots to them. We performed counter-balancing on the order that visualizations are shown to the user, and on the questionnaires asked for each visualization. We discuss the exact order and split of questionnaires here.

A.1 Order of Visualization

For each user, we show both scatterplot and bar-line chart, one after the other. In order to avoid learning effects of the first visualization on the second, we changed the order of showing the visualizations. Accordingly, out of 23 users, for the first 11 users, we showed bar-line chart first and scatterplot second. For the next 12 users, we reversed the order and showed scatterplot first and bar-line chart second.

A.2 Order of questionnaires

For each visualization, we ask users a set of questions that test the user's understanding of the particular visualization. More specifically, for each user, we ask them to identify:

1. globally first and second most consumed genre
2. the user's first and second most consumed genre
3. the user's first and second highest blind-spot genre
4. artist hotness value of these genres (i.e., the user's most consumed and blind-spot genres)

In addition to these 12 questions, for each question we also ask users to provide their confidence in their answer to that question (1-5 rating scale). Hence there were a total of 24 questions that the users were asked to answer. However, to avoid learning effect, we decided to split these 24 questions into two sets (Table A.1), and counter-balance on the order of placement of these sets for each visualization. The questions were split in such a way that for each visualization, the user answers at least one question about the global consumption pattern, the user's consumption pattern and the user's blind-spots.

TABLE A.1: Questionnaire split - 12 questions per set, i.e., the given 6 questions per set and confidence values for each question. (x3) means the question is repeated thrice.

Set	**Questions**
Set 1	Globally first highest genre
	User's first highest genre
	User's second highest blind-spot
	Artist hotness values for the above genres (x3)
Set 2	Globally second highest genre
	User's second highest genre
	User's first highest blind-spot
	Artist hotness values for the above genres (x3)

Counter-balancing was done on the order of these sets based on the user's ID, which was assigned to each user during registration. The user ID is a unique integer value assigned for each user, and counter-balancing was done in such a way that users with even ID get Set 1 for scatterplot and set 2 for the bar-line chart, and users with odd ID get set 2 for scatterplot and set 1 for the bar-line chart. This technique ensured that in the end, the different combinations of question sets and visualizations were distributed equally among the users. Table A.2 shows the number of users for each combination of the order of visualizations and the order of appearance of question sets. As we see from the table, the split is even with almost equal users in each for each combination.

TABLE A.2: Number of users for each combination of question-sets and order of visualization

	Order of Visualizations		
Questionnaire set		Bar-line chart first	Scatterplot first
	Set 1	6	6
	Set 2	5	6

Bibliography

[1] Panagiotis Adamopoulos and Alexander Tuzhilin. “On unexpectedness in recommender systems: Or how to better expect the unexpected”. In: (2014).

[2] Gediminas Adomavicius and Youngok Kwon. “Improving aggregate recommendation diversity using ranking-based techniques”. In: (2012).

[3] Gediminas Adomavicius and YoungOk Kwon. “Maximizing aggregate recommendation diversity: A graph-theoretic approach”. In: (2011).

[4] Srikant R. Agrawal R. “Fast algorithms for mining association rules.” In: (1997).

[5] Bart P. Knijnenburg et al. “Explaining the user experience of recommender systems”. In: (2012).

[6] Cai-Nicolas Ziegler et al. “Improving recommendation lists through topic diversification”. In: (2005).

[7] Makoto Nakatsuji et al. “Classical music for rock fans?: novel recommendations for expanding user interests”. In: (2010).

[8] Martijn C. Willemsen et al. “Using latent features diversification to reduce choice difficulty in recommendation lists”. In: (2011).

[9] Michael D. Ekstrand et al. “User perception of differences in recommender algorithms”. In: (2014).

[10] Soude Fazeli et al. “User-centric Evaluation of Recommender Systems in Social Learning Platforms: Accuracy is Just the Tip of the Iceberg”. In: (2017).

[11] Tao Zhou et al. “Solving the apparent diversity-accuracy dilemma of recommender systems”. In: (2010).

[12] Yuan Cao Zhang et al. “Auralist: introducing serendipity into music recommendation”. In: (2012).

[13] Pek Van Andel. “Anatomy of the unsought finding Serendipity: Origin, history, domains, traditions, appearances, patterns and programmability”. In: (1994).

[14] Ricardo Baeza-Yates and Berthier Ribeiro-Neto. “Modern information retrieval”. In: (1999).

[15] Adamic LA Bakshy E Messing S. “Exposure to ideologically diverse news and opinion on Facebook”. In: (2015).

[16] Whitman B Lamere P. Bertin-Mahieux T Ellis DP. “The Million Song Dataset”. In: (2011).

[17] López C Parra D Jeng Wy Brusilovsky P Oh JS. “Linking information and people in a social system for academic conferences”. In: (2017).

[18] Borgelt C. “Keeping things simple: finding frequent item sets by recursive elimination”. In: (2005).

[19] Borgelt C. “Simple algorithms for frequent item set mining”. In: (2010).

[20] Jos´e Campos and Antonio Dias de Figueiredo. “Searching the unsearchable: Inducing serendipitous insights”. In: (2001).

[21] Goldstein J Carbonell J. “The use of MMR, diversity-based reranking for reordering documents and producing summaries”. In: (1998).

[22] Armelle Brun Castagnos Sylvain and Anne Boyer. “When Diversity Is Needed... But Not Expected!” In: (2013).

[23] Òscar Celma Herrada. “Music recommendation and discovery in the long tail”. In: (2009).

[24] Shuo Chang, F Maxwell Harper, and Loren Gilbert Terveen. “Crowd-Based Personalized Natural Language Explanations for Recommendations”. In: *Proceedings of the 10th ACM Conference on Recommender Systems*. ACM. 2016, pp. 175–182.

[25] *Data-Driven Documents*. https://d3js.org/.

[26] Koren Y Weimer M Dror G Koenigstein N. “The yahoo! music dataset and kdd-cup’11”. In: (2011).

[27] Pariser E. “The filter bubble: What the Internet is hiding from you”. In: (2011).

[28] *Echo Nest Analyze, API*. http://static.echonest.com/enspex/.

[29] Ryokai K Goldberg K. Faridani S Bitton E. “Opinion space: a scalable tool for browsing online comments”. In: (2010).

[30] *Flask: web development, one drop at a time*. http://flask.pocoo.org/.

[31] Carla Delgado-Battenfeld Ge Mouzhi and Dietmar Jannach. “Beyond accuracy: evaluating recommender systems by coverage and serendipity”. In: (2010).

[32] Fatih Gedikli Ge Mouzhi and Dietmar Jannach. “Placing high diversity items in Top-N Recommendtion Lists”. In: (2011).

[33] Neil Hunt Gomez-Uribe Carlos A. “The netflix recommender system: Algorithms, business value, and innovation”. In: (2016).

[34] Schreiber H. “Improving Genre Annotations for the Million Song Dataset”. In: (2015).

[35] Yin Y. Han J Pei J. “Mining frequent patterns without candidate generation.” In: (2000).

[36] Jonathan L. et al. Herlocker. “Evaluating collaborative filtering recommender systems”. In: (2004).

[37] Riedl J. Herlocker JL Konstan JA. “Explaining collaborative filtering recommendations”. In: (2000).

[38] Rong Hu and Pearl Pu. “Enhancing recommendation diversity with organization interfaces”. In: (2011).

[39] Pu P Hu R. “Helping Users Perceive Recommendation Diversity.” In: (2011).

[40] Li D Hu Y. “Evaluation on Feature Importance for Favorite Song Detection”. In: (2013).

[41] Ogihara M. Hu Y. “Genre classification for million song dataset using confidence-based classifiers combination”. In: (2012).

[42] Neil J. Hurley. “Personalised ranking with diversity”. In: (2013).

[43] Masayuki et al Ishikawa. “Long tail recommender utilizing information diffusion theory”. In: (2008).

[44] Cohen J. “Statistical power analysis for the behavioral sciences. 2nd”. In: (1988).

[45] Qiang Guo Jian-Guo Liu Kerui Shi. “Solving the accuracy-diversity dilemma via directed random walks”. In: (2012).

[46] Verbert K Jin Y Tintarev N. “Effects of individual traits on diversity-aware music recommender user interfaces.” In: (2018).

[47] Derek Bridge Kaminskas Marius. "Diversity, Serendipity, Novelty, and Coverage: A Survey and Empirical Analysis of Beyond-Accuracy Objectives in Recommender Systems". In: (2016).

[48] Höllerer T O'Donovan J Kang B Tintarev N. "What am I not seeing? An interactive approach to social content discovery in microblogs". In: (2016).

[49] Wagener T Kelleher C. "Ten guidelines for effective data visualization in scientific publications." In: (2011).

[50] Christos Faloutsos Kensuke Onuma Hanghang Tong. "TANGENT: A novel, "surprise me," recommendation algorithm". In: (2009).

[51] Bart P. Knijnenburg et al. "Inspectability and Control in Social Recommenders". In: *Conference on Recommender Systems*. RecSys '12. Dublin, Ireland, 2012, pp. 43–50. ISBN: 978-1-4503-1270-7. DOI: 10.1145/2365952.2365966. URL: http://doi.acm.org/10.1145/2365952.2365966.

[52] Loren Terveen Joseph A. Konstan Paul Schrater Komal Kapoor Vikas Kumar. "I like to explore sometimes: Adapting to dynamic user novelty preferences". In: (2015).

[53] Niemela F Kreitz G. "Spotify–large scale, low latency, P2P music-on-demand streaming". In: (2010).

[54] Freelon D Borning A Bennett L Kriplean T Morgan J. "Supporting reflective public thought with considerit". In: (2012).

[55] Morgan J Borning A Ko A Kriplean T Toomim M. "Is this what you meant?: promoting listening on the web with reflect". In: (2012).

[56] Pasquale Lops Giovanni Semeraro Michele Filannino Piero Molino Leo Iaquinta Marco de Gemmis. "Introducing serendipity in a content-based recommender system". In: (2008).

[57] Cass R. Sunstein Leonard TC. Richard H. Thaler. "Nudge: Improving decisions about health, wealth, and happiness". In: 2008.

[58] Li Q Li T Ogihara M. "A comparative study on content-based music genre classification". In: (2003).

[59] Ellis DP Lanckriet GR McFee B Bertin-Mahieux T. "The million song dataset challenge". In: (2012).

[60] Stephanie Y. Lee Paul Resnick Munson Sean A. "Encouraging Reading of Diverse Political Viewpoints with a Browser Widget." In: (2013).

[61] Julita Vassileva Nagulendra Sayooran. "Understanding and controlling the filter bubble through interactive visualization: A user study". In: (2014).

[62] Smyth B Nava T Rostami S. "Knowing the unknown: visualising consumption blind-spots in recommender system". In: (2018).

[63] Robbins NB. "Creating more effective graphs". In: (2012).

[64] Lamere P. "Social tagging and music information retrieval." In: (2008).

[65] Vyas OP Pramod S. "Survey on frequent item set mining algorithms". In: (2010).

[66] Li Chen Pu Pearl. "Trust building with explanation interfaces". In: (2006).

[67] Li Chen Pu Pearl and Rong Hu. "A user-centric evaluation framework for recommender systems". In: (2011).

[68] Pearl Pu et al. "Usability guidelines for product recommenders based on example critiquing research". In: *Recommender Systems Handbook*. Springer, 2011, pp. 511–545.

[69] *Pymining*. https://github.com/bartdag/pymining/tree/master/pymining.

[70] Veloso A Ziviani N Ribeiro MT Lacerda A. "Pareto-efficient hybridization for multi-objective recommender systems". In: (2012).

[71] Lei Shi. "Trading-off among accuracy, similarity, diversity, and long-tail: A graph-based recommendation approach". In: (2013).

[72] Wang J Larson M Hanjalic A Shi Y Zhao X. "Adaptive diversification of recommendation results via latent factor portfolio". In: (2012).

[73] McClave P Smyth B. "Similarity vs. diversity. Case-Based Reasoning Research and Development". In: (2001).

[74] Chen K Yu Y Su R Yin LA. "Set-oriented personalized ranking for diversified top-n recommendation". In: (2013).

[75] *The Last.fm Dataset*. https://labrosa.ee.columbia.edu/millionsong/lastfm.

[76] Judith Masthoff Tintarev Nava. "Explaining recommendations: Design and evaluation". In: (2015).

[77] Höllerer T O'Donovan J Tintarev N Kang B. "Inspection Mechanisms for Community-based Content Discovery in Microblogs". In: (2015).

[78] Chun-Hua Tsai and Peter Brusilovsky. "Leveraging interfaces to improve recommendation diversity." In: (2017).

[79] Peter Brusilovsky Tsai Chun-Hua. "Providing Control and Transparency in a Social Recommender System for Academic Conferences". In: (2017).

[80] Brusilovsky P Tsai CH. "Enhancing Recommendation Diversity Through a Dual Recommendation Interface". In: (2017).

[81] Cook P. Tzanetakis G. "Musical genre classification of audio signals". In: (2002).

[82] Sa'ul Vargas and Pablo Castells. "Improving sales diversity by recommending users to items". In: (2014).

[83] Karatzoglou A Castells P Vargas S Baltrunas L. "Coverage, redundancy and size-awareness in genre diversity for recommender systems". In: (2014).

[84] Vallet D Vargas S Castells P. "Intent-oriented diversity in recommender systems". In: (2011).

[85] Zhu J Wang J. "Portfolio theory of information retrieval". In: (2009).

[86] Xinyu et al Xing. "Exposing inconsistent web search results with bobble". In: (2014).

[87] Laks VS Lakshmanan Yu Cong and Sihem Amer-Yahia. "Recommendation diversification using explanations". In: (2009).

[88] Ogihara M Li W Zaki MJ Parthasarathy S. "New Algorithms for Fast Discovery of Association Rules". In: (1997).

[89] Hurley N Zhang M. "Avoiding monotony: improving the diversity of recommendation lists". In: (2008).