

Personalized Explanations for Hybrid Recommender Systems

Pigi Kouki
UC Santa Cruz
pkouki@soe.ucsc.edu

James Schaffer
Army Research Lab
james.a.schaffer20.civ@mail.mil

Jay Pujara
University of Southern California
jay@cs.umd.edu

John O'Donovan
UC Santa Barbara
jod@cs.ucsb.edu

Lise Getoor
UC Santa Cruz
getoor@soe.ucsc.edu

ABSTRACT

Recommender systems have become pervasive on the web, shaping the way users see information and thus the decisions they make. As these systems get more complex, there is a growing need for transparency. In this paper, we study the problem of generating and visualizing personalized explanations for *hybrid* recommender systems, which incorporate many different data sources. We build upon a hybrid probabilistic graphical model and develop an approach to generate real-time recommendations along with personalized explanations. To study the benefits of explanations for hybrid recommender systems, we conduct a crowd-sourced user study where our system generates personalized recommendations and explanations for real users of the last.fm music platform. We experiment with 1) different explanation styles (e.g., user-based, item-based), 2) manipulating the number of explanation styles presented, and 3) manipulating the presentation format (e.g., textual vs. visual). We apply a mixed model statistical analysis to consider user personality traits as a control variable and demonstrate the usefulness of our approach in creating personalized hybrid explanations with different style, number, and format.

CCS CONCEPTS

• **Information systems** → **Decision support systems**; *Collaborative filtering*; • **Human-centered computing** → *Social networking sites*; *Empirical studies in visualization*.

KEYWORDS

Explainable artificial intelligence, explainable intelligent user interfaces, hybrid recommender systems, explainable recommender systems

ACM Reference Format:

Pigi Kouki, James Schaffer, Jay Pujara, John O'Donovan, and Lise Getoor. 2019. Personalized Explanations for Hybrid Recommender Systems. In *Proceedings of 24th International Conference on Intelligent User Interfaces (IUI '19)*. ACM, New York, NY, USA, 12 pages. <https://doi.org/10.1145/3301275.3302306>

ACM acknowledges that this contribution was authored or co-authored by an employee, contractor, or affiliate of the United States government. As such, the United States government retains a nonexclusive, royalty-free right to publish or reproduce this article, or to allow others to do so, for government purposes only.

IUI '19, March 17–20, 2019, Marina del Rey, CA, USA

© 2019 Association for Computing Machinery.

ACM ISBN 978-1-4503-6272-6/19/03.

<https://doi.org/10.1145/3301275.3302306>

1 INTRODUCTION

Recommender systems have become an essential tool for consumers to navigate the vast number of options for content and products. As recommendations have become central to combating information overload and shaping decisions, users increasingly demand convincing explanations to help them understand why particular recommendations are made [7, 25]. The increasing complexity of recommender systems has fueled a corresponding need for explanations to evolve to capture the richness of information used to make a recommendation. However, individual differences also require catering to users' preferences for the explanations they find most persuasive.

Most of the research on explaining recommendations (e.g., [15, 6, 33, 40, 9]) has focused on explanations that involve a single source of data. Typically, explanations from single-source recommenders come in a *single style*, e.g., a content-based recommender system produces content-based explanations. However, modern recommendation frameworks combine information from diverse sources such as social connections, collaborative filtering (CF) approaches, and item metadata, an approach known as hybrid recommendation. A recent survey [28] of different single-style explanation approaches has concluded that hybrid explanations, which combine multiple styles such as user-based and item-based, are more effective than non-hybrid counterparts. Despite these findings, there has been no comprehensive study to determine the best methods for presenting personalized recommendations with explanations that are generated for hybrid systems.

In this paper, we implemented a hybrid recommender system that can incorporate a variety of recommendation approaches (e.g., content-based and CF). Using this hybrid system, we generated hybrid explanations which consist of many *styles*, where each style is associated with a recommendation algorithm. To better understand how to improve explanation persuasiveness using multiple explanation styles, we conducted a large personalized user study on a real system from the last.fm music domain. We experimentally manipulated the explanation styles produced (e.g., social-based or content-based), the number of explanation styles presented (e.g., three or four explanations), and the visual format of the explanations (e.g., textual or visual). Recent research [5] has indicated that there may be a relationship between a person's personality and the type of explanation that is most persuasive, so we also conducted an exploratory analysis for hybrid explanations. To the best of our knowledge, our work is the first study of the effect of such variables on *personalized hybrid explanations*.

In this work, we extended an existing hybrid recommender system [21], to produce real-time recommendations while incorporating a variety of information sources. Moreover, we built a real-time data collection system for acquiring a user’s history, social connections, tags, and popularity statistics from the last.fm social media site. We used these signals to create a hybrid model that incorporates user-user and item-item similarities using CF, content, social, and popularity information. Next, we implemented a translation system that generates customized explanations from the output of the hybrid system in real time. More explicitly, we converted each signal to a different explanation style. The proposed model supports seven different signals which are translated to seven different explanation styles, including user-based, item-based, content, social, and item popularity. Although the explanation styles are fixed, their content is *personalized* to each user’s data and personality. For example, the social style explanation would be of the form “We recommend z because your friend x likes z ” but the values of z and x will be tailored based on each user’s data, and potentially used less frequently for introverted users. Table 1 shows an example of a personalized recommendation along with the personalized explanations generated by our framework for a particular user.

In the second part of our work, we generated real-time recommendations along with personalized explanations for users of the last.fm music platform. We conducted a crowd-sourced user study ($N = 198$) using Amazon Mechanical Turk (AMT), recruiting users with active last.fm accounts. Our first goal was to study whether different explanation styles result in different levels of persuasiveness. For example, are social explanations perceived as more persuasive than popularity ones? Inspired by Berkovsky et al. [5], we also considered the personality traits of the users as a control variable. We find interesting patterns between explanation persuasiveness of particular styles and personality traits which we analyze in our results. Second, we studied whether the number of the explanation styles can affect the persuasiveness of the explanation. For example, is a user convinced when they are provided with three explanation styles but overwhelmed when the number of provided styles increases to six? Third, we experiment with a variety of formats that we can present hybrid explanations to the users. For example, do users prefer to see explanations in visual or textual formats?

Finally, we argue that explainable recommendations should not require a trade-off in recommendation quality. To this end, in an offline setting, we compared the accuracy of the hybrid model with popular collaborative filtering algorithms, such as item-based and matrix factorization. We show that our model improves the accuracy of recommendations over baseline collaborative filtering methods while at the same time being able to provide explanations.

Our contributions are: (1) a hybrid recommender system that can provide real-time recommendations with up to seven different explanation styles in various formats (such as textual and visual), (2) insights regarding the most persuasive styles, the ideal number of explanation styles, and the most persuasive presentation formats as a result of a user study, and (3) insights regarding the effect of different personality traits in persuasiveness of explanations (as a result of the same study).

In the next sections, we discuss related work, present our hybrid recommendation framework, and define our research questions. We then describe the user study and present the results. Afterwards,

Explanation Style	We recommend U_2 because:
(I) User-based	User <i>Aren</i> with whom you share similar tastes in artists, listens to U_2 .
(II) Item-based	(a) People who listen to your profile item <i>AC/DC</i> also listen to U_2 . (b) Last.fm’s data indicates that U_2 is similar to <i>Coldplay</i> that is in your profile.
(III) Content	(a) U_2 has similar tags as <i>Beatles</i> that is in your profile. (b) U_2 is tagged with <i>rock</i> that is in your profile.
(IV) Social	Your friend <i>Cindy</i> likes U_2 .
(V) Item popularity	U_2 is a very popular in the last.fm database with 3.5 million listeners and 94 million playcounts.

Table 1: An example of a hybrid explanation for a single artist (U_2). Multiple styles are generated from the hybrid model (the first four are personalized, while the fifth one is non-personalized).

we evaluate the hybrid model and a set of baseline recommendation algorithms in an offline setting. Finally, we summarize our findings, point out limitations, and discuss our future research plans.

2 RELATED WORK

There is an emerging body of work on explanations for recommenders systems (whether single-source or hybrid). We review the most representative work and discuss the basic differences between our work and the state-of-the-art.

Single-style explanations. Herlocker et al. [15] showed that certain explanation and presentation styles can increase a recommender system’s effectiveness in convincing users to make a purchase. Bilgic and Mooney [6] compared single-style explanations that use content-based keywords, item-based CF, or prior rating history. Vig et al. [40] showed that explanations using tags improve effectiveness. Tintarev and Masthoff [35] found that, despite improving user satisfaction, personalization can reduce the effectiveness of content-based explanations. Gedikli et al. [13] studied the effect of ten single-style explanations (personalized or not) for the following dimensions: efficiency, effectiveness, persuasiveness, transparency, and satisfaction. The authors showed that non personalized content-based explanations in the form of tags increase user-perceived transparency and satisfaction. In a music recommendation setting, Oramas et al. [27] found that the effectiveness of an explanation depends on the familiarity with recommender systems and the music education of the users. PeerChooser [26] is a system designed to present user-based CF explanations through an interactive graphical interface in the form of concentric circles. Berkovsky et al. [5] studied the effect of three different explanation styles (item-based, average rating, and popularity-based) on user trust considering the user personality traits as a control variable. In our work, we study whether varying the explanation style or number of explanations changes the persuasiveness of explanations when controlling for different personality traits.

Hybrid explanations. Most research on hybrid explanations has focused on proposing graphical interfaces to visualize the different explanation styles. TalkExplorer [39] combines content, tag, and social-based filtering techniques to provide an interactive interface in the form of clustermaps. SetFusion [29] builds on TalkExplorer

and replaces clustermaps with Venn diagrams showing improved user experience. TasteWeights [7] builds an interactive hybrid recommender system that combines social, content, and expert information. The framework shows the reasoning behind the recommendations in the form of pathways among columns. Nguyen et al. [23] aimed to reduce the noise in user ratings by proposing interfaces that support explanations in the form of tags, exemplars, and their combination. Symeonidis et al. [33] combined content-based filtering and rating history to generate natural language explanations which are all of the same type. Sato et al. [31] proposed to use contexts for explanations (e.g., “accompanying persons”) and show that combining context with other styles of explanations (e.g., demographic) improves persuasiveness and usefulness. Recently, Andjelkovic et al. [2] implement MoodPlay, a hybrid music recommender system that combines content and mood-based filtering. The system provides interaction, control, and explanations. Explanations are provided using links to the last.fm music profile of the recommended items and using circles representing mood. Finally, in our previous work [22] we manually generated hybrid explanations in a restaurant recommendation setting. We conducted a synthetic user study where users evaluated non-personalized explanations that were manually produced. In this paper, we implement a hybrid recommender system which automatically generates recommendations together with explanations. We use this system to generate personalized, real-time recommendations with explanations for active users of a music platform. In our experiments we analyze both the recommendation quality and the explanation persuasiveness by varying several different variables.

Nunes and Jannach [25] reviewed the literature on explanations in decision-support systems. The authors distinguished between variables such as the length of the explanation, its vocabulary, and the presentation of the explanation. The conclusion was that additional studies are necessary to assess the impact of these variables. One of the goals of our work is to determine whether the explanation length and its presentation affect user satisfaction. In another work, Friedrich and Zanker [11] proposed a taxonomy that categorizes different explainable recommender systems. The authors argue that future research should create new kinds of information, interaction, and presentation styles and also analyze how, and under what conditions, these will affect different explanation objectives. To this end, we offer seven different explanation styles and study their effect on persuasiveness when taking different variables into account.

To summarize, our work differs from prior work in the following ways. First, existing work on explanations either does not involve a recommendation algorithm [5], or uses a baseline recommender [7, 27, 13]. As a result, comparing the accuracy of the recommendations of these approaches with other algorithms is prohibitive. Here, we generate explanations from an existing hybrid recommender system and show in an offline setting that this system can outperform popular recommendation methods, such as matrix factorization. Second, our work considers seven different explanation styles, while most of prior work considers up to three explanation styles. Third, to the best of our knowledge, our user study is the first one analyzing the effect of different personalized explanation styles, their number, and format on the persuasiveness of explanations. Finally,

our study is the first that considers the user personality traits as a control variable.

3 EXPLAINABLE HYBRID RECOMMENDER

In this section, we describe how we use a hybrid recommender system to generate explainable recommendations. We first introduce a modeling framework which constructs a probabilistic model using a set of rules capturing relational dependencies. Next, we describe how we use this framework to implement a music recommender system. Finally, we discuss how we transform the model’s probabilistic factors to explanations capturing the different recommender signal types.

3.1 Probabilistic model for hybrid recommendations

Recently, Kouki et al. [21] introduced HyPER, a hybrid recommender system that uses the probabilistic soft logic (PSL) framework [3]. PSL is a generic probabilistic programming language that allows for models to be defined through a set of template rules in first-order logic syntax. The rules can incorporate a wide range of signals, such as user-user and item-item similarity measures, content, and social information. HyPER uses rules together with the input data to perform inference and define a probability distribution over the recommended items, capturing the likelihood that a given user will like a given item. HyPER provides a generic and extensible recommendation framework with the ability to incorporate other sources of information available in different application scenarios. Here, we focus on music recommendations. We use the rules proposed in HyPER and add several rules to leverage additional information available in our music dataset.

3.2 Hybrid music recommender model

We propose a hybrid music-recommender system which consists of the following rules:¹

$$\text{SIMUSERS}_{CF}(u_1, u_2) \wedge \text{LISTENS}(u_1, a) \Rightarrow \text{LISTENS}(u_2, a) \quad (1)$$

$$\text{SIMARTISTS}_{CF}(a_1, a_2) \wedge \text{LISTENS}(u, a_1) \Rightarrow \text{LISTENS}(u, a_2) \quad (2)$$

$$\text{SIMARTISTS}_{last.fm}(a_1, a_2) \wedge \text{LISTENS}(u, a_1) \Rightarrow \text{LISTENS}(u, a_2) \quad (3)$$

$$\text{SIMARTISTS}_{content}(a_1, a_2) \wedge \text{LISTENS}(u, a_1) \Rightarrow \text{LISTENS}(u, a_2) \quad (4)$$

$$\text{HASTAG}(a_1, t) \wedge \text{HASTAG}(a_2, t) \wedge \text{LISTENS}(u, a_1) \Rightarrow \text{LISTENS}(u, a_2) \quad (5)$$

$$\text{SIMFRIENDS}(u_1, u_2) \wedge \text{LISTENS}(u_1, a) \Rightarrow \text{LISTENS}(u_2, a) \quad (6)$$

$$\text{POPULARARTIST}(a) \Rightarrow \text{LISTENS}(u, a) \quad (7)$$

$$\neg \text{LISTENS}(u, a) \quad (8)$$

Rule 1 captures the intuition that similar users like similar artists. An atom such as $\text{LISTENS}(u_2, a)$ represents the probability that user u_2 will listen to artist a and takes values in the interval $[0, 1]$. Higher atom values indicate a higher probability that the given user will listen to the given artist. Atom $\text{SIMUSERS}_{CF}(u_1, u_2)$ is binary, with value 1 iff u_1 is one of the k -nearest neighbors of u_2 . We compute user similarities using CF information (indicated by the CF subscript). More specifically, we compute similar users using the Jaccard and cosine similarities. Jaccard similarity between two users is computed using the set of common artists they have listened to:

¹The code is available here: <https://github.com/pkouki/iui19>.

$$J(u_1, u_2) = \frac{|Artists(u_1) \cap Artists(u_2)|}{|Artists(u_1) \cup Artists(u_2)|} \quad (9)$$

where $Artists(u_i)$ denotes the set of artists that user u_i has listened to. Cosine similarity is computed using vectors containing the number of times a user listened to each artist:

$$Cos(u_1, u_2) = \frac{\sum_{a \in \mathcal{A}} ListenCount(u_1, a) ListenCount(u_2, a)}{\sqrt{\sum_{a \in \mathcal{A}} ListenCount(u_1, a)^2} \sqrt{\sum_{a \in \mathcal{A}} ListenCount(u_2, a)^2}}$$

where \mathcal{A} is the set of all artists, and $ListenCount(u_i, a)$ indicates how many times user u_i has listened to artist a . The number of similar users is typically set to between 20 and 50 in the literature [24], and so for each user we use the 20 most similar neighbors. This limit applies to all similarities that we describe in the rest of this section.

Rule 2 captures the intuition that a user listens to similar artists. Artist similarity is computed using CF information by computing the Jaccard similarity of the sets of users who have listened to each artist (we follow the same logic as in formula 9). Rule 3 is similar with the difference that we use last.fm’s artist similarity, which is a proprietary score using CF and tag information. Rule 4 captures the intuition that users are likely to listen to artists with similar content. We measure content similarity using tags associated with each artist and compute the Jaccard similarity between the tag sets of two artists (again we follow the same logic as in formula 9). Rule 5 is a simpler version of Rule 4 and captures the intuition that a user will likely listen to two artists sharing the same tag.

Rule 6 captures the intuition that friends may listen to the same artists. Rule 7 captures the intuition that a user will likely listen to a popular artist from the last.fm database. Every music website offers a large number of artists, however, in the general case, a user listens only to a very small portion of the artists provided. To model our general belief that a user will likely not listen to an artist we introduce the Rule 8. Finally, we note that the model can incorporate new information sources by adding additional first-order rules.

3.3 Generating recommendations with personalized explanations

The rules used in HyPER specify probabilistic dependencies between variables and evidence. After encoding all available information, e.g., similarities and observed user-item likes, the next step is to use our model for predicting unobserved user-item likes. The process of combining the model with data and instantiating a set of propositions is referred to as *grounding*. Each ground rule is translated into a hinge-loss potential function. The set of ground rules defines a probabilistic graphical model and in particular a Markov random field. Performing inference over this model generates predictions for unobserved user-artist pairs, captured by the LISTENS predicate. In other words, we find the most probable assignment to the unobserved variables (LISTENS) by performing joint inference over interdependent variables. After the inference completes for a user u , we select the LISTEN(u, a) that scored in the top k positions. For each of the top k LISTENS(u, a), we use the groundings generated during inference to create personalized explanations of the following styles:

- **User-based**, with explanations similar to the example of Table 1 (I) using the groundings of Rule 1.
- **Item-based CF** and **item-based last.fm**, with explanations similar to Table 1 (II-a, II-b), using the groundings of Rules 2 and 3 respectively.
- **Content-based Jaccard** and **content-based tags**, with explanations similar to Table 1 (III-a, III-b) using the groundings of Rules 4 and 5 respectively.
- **Social-based**, with explanations similar to Table 1 (IV) using the groundings of Rule 6.
- **Popularity-based**, with explanations similar to Table 1 (V) using the groundings of Rule 7.

As an example, let’s assume that for user *Jen*, the predicted value of the unobserved variable LISTENS(*Jen*, *U2*) has the highest value among all other predicted values and, during inference, the following ground rules associated with LISTENS(*Jen*, *U2*) were generated:

$$\begin{aligned} SIMUSERS_{CF}(Jen, Aren) \wedge LISTENS(Aren, U2) &\Rightarrow LISTENS(Jen, U2) \\ SIMARTISTS_{CF}(U2, ACDC) \wedge LISTENS(Jen, ACDC) &\Rightarrow LISTENS(Jen, U2) \\ SIMARTISTS_{last.fm}(U2, Coldplay) \wedge LISTENS(Jen, Coldplay) &\Rightarrow LISTENS(Jen, U2) \\ SIMARTISTS_{content}(U2, Beatles) \wedge LISTENS(Jen, Beatles) &\Rightarrow LISTENS(Jen, U2) \\ HAS_{TAG}(U2, Rock) \wedge HAS_{TAG}(Slayer, Rock) \wedge LISTENS(Jen, Slayer) &\Rightarrow LISTENS(Jen, U2) \\ SIMFRIENDS(Jen, Cindy) \wedge LISTENS(Cindy, U2) &\Rightarrow LISTENS(Jen, U2) \\ POPULARARTIST(U2) &\Rightarrow LISTENS(Jen, U2) \end{aligned}$$

In order to generate explanations from the ground rules, we developed a translation system that takes as input the groundings and outputs sentences in natural language. Table 1 shows the natural language explanations generated by the ground rules shown in this specific example. Note that not all explanation styles will be present for every recommendations, resulting in some missing styles. In our empirical study, we found that most recommendations had all styles. Finally, we underscore the flexibility of our approach. If new rules are added to the model, a similar process can be used to generate explanation styles corresponding to those rules.

4 LAST.FM DATASET

We evaluated our system on music recommendations for the last.fm website. We chose this platform because: i) it provides an API² offering convenient access to music data and ii) it contains a wide range of information that can be exploited by the hybrid model: user-artist interactions, user friendships, content information for artists (i.e., tags), and popular artists in the database. Last.fm exposes two main API types, *Users* and *Artists*. The *User* API provides access to the user’s top artists by listening frequency and the user’s friends. The *Artist* API provides access to similar artists to a given artist (computed using CF and tag information) and the top user-provided tags for an artist. Last.fm also offers general top-chart methods returning information such as the k artists with the highest number of listeners and the k tags appearing the highest number of times across last.fm’s database. To integrate with last.fm’s API we built a crawler using pylast³ that allows us to collect information for each user in our study in real time.

²<https://www.last.fm/api>

³<https://github.com/pylast/pylast>

5 RESEARCH QUESTIONS

Our work addresses the following basic research questions about recommendations and explanations for a hybrid system:

1. How does explanation persuasiveness vary with different explanation styles? An explanation from a hybrid recommender system usually contains several different styles, such as user-based and social-based. Our goal is to study whether varying these styles changes the persuasiveness of an explanation. Additionally, following prior work [36] showing that personality strongly correlates with users' characteristics used by recommender systems (e.g., music preferences), we study whether there are differences in preferred explanation style when we use personality traits as a predictor of explanation preferences. Our hypothesis is that users with specific personal characteristics will be persuaded to different degrees by different explanation styles. For example, an extrovert may be receptive to social style explanations, while an introvert may prefer item-based style explanations.

2. What is the ideal number of explanation styles? One pitfall in explanatory systems is information overload. We identify the inflection point in terms of the number of styles at which users lose interest. We vary the number of different explanation styles presented to the user for each recommendation. Our hypothesis is that different number of explanation styles will result in different persuasiveness levels. Our goal is to determine the optimal number of explanation styles that balance information overload and persuasiveness. We additionally study whether there is any difference when we take the user personality traits into account.

3. How does the explanation format affect user experience? Prior work on non-personalized explanations [22] showed that user experience is affected by the format of the explanations, i.e., users prefer simple visual formats over complex ones. Based on these results, we study the effect of textual and simple visual formats (Venn diagrams and cluster dendrograms) in personalized explanations. Our hypothesis is that different visual formats will result in different levels of user experience.

6 USER-STUDY DESIGN

We used the AMT platform to recruit active last.fm participants for our user study. In this section, we describe our study from the point of view of one participant. The study was divided in two phases. In the first phase, we asked the participant to fill in a pre-study questionnaire (this is the same for all participants). At the same time, we crawled this participant's music data and ran the HyPER model to generate recommendations with explanations. In the second phase of the study, we showed the produced personalized recommendations with explanations to the participant and ask a set of questions by following a methodology similar to Knijnenburg et al.'s [20]. Note that the recommended artists and the actual explanations are personalized to each participant, while the questions for evaluating recommendations and explanations are the same for all participants.

6.1 First phase: pre-study and generation of recommendations and explanations

In this first phase of the study, we informed the participant that, in order to participate, she needed to have a last.fm account with

Ease-of-Satisfaction ($\alpha = 0.89$)	R^2	$Est.$
I think I will trust the artists recommendations given in this task.	0.68	0.93
I think I will be satisfied with the artists recommendations given in this task.	0.89	1.11
I think the artist recommendations in this task will be accurate.	0.67	1.01
Visualization Familiarity ($\alpha = 0.92$)	R^2	$Est.$
I am competent when it comes to graphing and tabulating data.	0.75	1.44
I frequently tabulate data with computer software.	0.71	1.46
I have graphed a lot of data in the past.	0.78	1.52
I frequently analyze data visualizations.	0.68	1.46
Personality - I see myself as...	Trait	
Extroverted, enthusiastic.	Extroversion	
Reserved, quiet.	Dependability	
Dependable, self-disciplined.	Dependability	
Disorganized, careless.	Dependability	
Open to new experiences, complex.	Openness	
Conventional, uncreative.	Openness	
Calm, emotionally stable.	Neuroticism	
Anxious, easily upset.	Neuroticism	
Sympathetic, warm.	Agreeableness	
Critical, quarrelsome.	Agreeableness	

Table 2: Pre-study questions asked to the participants. Factors (ease-of-satisfaction and visualization familiarity) are determined by participant responses to subjective questions via factor analysis, which was done in R lavaan using the semtools package. R^2 reports the fit of the item to the factor. $Est.$ is the estimated loading of the item to the factor. α is Cronbach's alpha.

at least ten artists in her profile. We prompted participants that were interested in the study but did not have a last.fm account to follow detailed instructions on how to create an account and enrich their profile with the prerequisites. After the participant provided her last.fm id, we checked whether it satisfied the study prerequisites. When the prerequisites were not met, we reminded the user of the requirements to participate in the study. Once the prerequisites were fulfilled, we directed the user to answer the pre-study questionnaire.

In the pre-study questionnaire, we asked the participant questions related to ease-of-satisfaction [32] and visualization familiarity. We additionally asked questions related to the five basic dimensions of personality, called the Big Five traits [36]. We adopted the abbreviated questionnaire by Gosling et al. [14] which is both brief and highly reliable. We report all the pre-study questions in the first column of Table 2. Responses were provided using a 7-point Likert scale from "totally disagree" to "totally agree". During the time that the participant answered the pre-study questions, we sequentially perform the following tasks in the background:

Crawl data: Using the last.fm API, we crawled the top 20 artists for this participant's profile. Next, for each of these artists, we crawled the top 20 tags and the top 20 most similar artists. For each similar artist, we crawled the top 20 tags. Next, we retrieved the top 20 friends of this participant along with their top 20 favorite artists.

Candidate-set creation: For each participant u of the study, we created a set of candidate artists \mathcal{A} . For each artist $a \in \mathcal{A}$, we generated an unobserved predicate $\text{LISTENS}(u, a)$. The HyPER model made predictions for **all** the unobserved $\text{LISTENS}(u, a)$ predicates. Last.fm contains a large number of artists whose popularity follows a power-law distribution, where many users listen to a few, popular artists and most artists are in a long tail, with few listeners. Since recommendations must be generated quickly during Phase 1, we applied selection criteria to reduce the number of candidate artists ($|\mathcal{A}|$), as is common in ranking tasks [1]. However, to ensure that the recommended artists were personalized to each participant’s tastes we created a user-specific candidate set consisting of three sets of artists: (i) the 20 most similar artists (based on the last.fm similarity) in the participant’s profile, (ii) the 20 top artists for each of the participant’s friends, and (iii) the top 1,000 artists in the last.fm database.

Compute similarities: We computed similarities used by the HyPER model for each participant. More specifically, we computed: (i) similarities between the participant and other last.fm users (used by Rule 1), (ii) similarities between the participant and her friends (used by Rule 6), and (iii) content and CF similarities for the artists in the participant’s candidate set (used by Rules 2 and 4). Similarity computations occurred while the participant was completing Phase 1, and thus these computations had to occur as quickly as possible. Computing user-user and artist-artist CF similarities is a very expensive operation [24] and generating similarities for all last.fm users and artists is impractical in a real-time user study. To mitigate this issue, we computed user and item similarities using CF information from a smaller subset of the last.fm dataset containing 1,475 users, 8,672 artists, and 28,639 user-artists pairs.

Run the HyPER model: In this step, we ran the process of grounding the rules, where we combined the model described in Section 3 with the evidence and instantiated a set of propositions. The evidence consisted of similarities computed in the previous step, user-item interactions, social connections, tags, and popularity statistics. After grounding, we ran inference to predict the probability that participant u will listen to artist a ($a \in \mathcal{A}$) (i.e., predict the values of the unobserved predicates $\text{LISTENS}(u, a)$). At the end of inference we picked the predictions for the predicates $\text{LISTENS}(u, a)$ that scored the highest.

Organize the explanations: To organize the explanations, we grouped multiple explanations of the same style together [22]. For example, if there were three groundings of the Rule 1 with similar users *Aren*, *Sonia*, and *Mary*, we grouped those into one single sentence: “Users *Aren*, *Sonia*, and *Mary*, with whom you share similar tastes in artists, like *U2*”. Since the number of groundings for each rule could be very large, it was not possible to show all the groundings of a rule. In this case, we used a threshold, $t = 3$, and showed at most 3 groundings of each rule. To select which t groundings to show, we pick the groundings that involve the highest similarity values.

6.2 Second phase: main study

After generating the top k recommendations and organizing them, the next step was to present them to the participant. We worked towards answering questions related to the participant preferences’

toward different styles, number, and format of the explanations. At the same time we controlled for the accuracy and novelty of the recommendations. To this end, we showed each participant three artists that ranked in the top three positions after running the HyPER framework and ask questions about the accuracy and novelty of the recommendations and questions related to the explanations provided with a focus on the persuasiveness aspect. We organized the study around the three questions discussed.

Task for research question 1: We showed the participant an artist profile for the highest ranked artist predicted by the HyPER model. The artist profile consisted of the artist’s name, an official picture, and a link to the artist’s last.fm page. We did not provide an explanation for the recommendation. We asked the participant to rate the accuracy and novelty of the recommendation using the questions of Table 3 (under “Perceived Accuracy” and “Perceived Novelty”). Next, we showed the same artist profile with only **one** explanation style (e.g., user-based) and asked the participant to respond to the question “How persuasive is this explanation?” using a 7-point Likert scale (from “not persuasive at all” to “very persuasive”). Next, we used a different explanation style (e.g., social) for the same artist profile and ask the same question. We repeated the process for all explanation styles generated by the HyPER framework. To avoid any order-related biases, we randomized the presentation order of different explanation styles. With this task we tested the following hypotheses:

- H_1 : Explanation style significantly correlates with perceived persuasiveness.
- H_2 : Personality significantly correlates with perceived persuasiveness.

Task for research question 2: We showed the participant an artist profile for the second highest ranked artist as predicted by the HyPER model. We did not provide an explanation and asked the same questions related to perceived accuracy and novelty. Next, we showed the participant all the explanation styles that were generated by the HyPER framework. We asked the participant to rank the explanation styles from the most persuasive to the least persuasive. We gave the participant the option to rank only the styles that are interesting and omit those they find uninteresting. Again, we randomized the initial order of the styles. Figure 1 shows an example of the ranking question. With this task we tested the following hypotheses:

- H_3 : People prefer to see the maximum number of explanation styles available.
- H_4 : Personality significantly correlates with preferred number of explanation styles.

Task for research question 3: We showed the participant an artist profile for the third highest ranked artist as predicted by the HyPER model. We did not provide an explanation and asked the same questions related to perceived accuracy and novelty. Next, we presented the same recommended artist with the same explanation styles using different formats (one textual and three visual). For each format we asked the participant to respond to a set of user experience (UXP) statements presented in Table 3 (under “Reception (UXP)”) using a 7-point Likert scale. To determine which visualizations to show, we used the results of a non-personalized crowd-sourced study [22], which showed that Venn diagrams significantly outperform more complex visualizations such as concentric circles and

Based on your last.fm profile, we recommend **Black Sabbath**.



For this recommendation, please consider the following explanations that are given below in the green boxes. Then, drag the explanations to rank them in order of persuasiveness, according to you. Please rank these items in order of persuasiveness to you. You do not have to rank all of the explanations, if some are not persuasive, please leave them in the lower box.

Move items here.

People who listen to your profile items *Metallica, Iron Maiden, Rainbow* also listen to to Black Sabbath.

The last.fm users with whom who share similar music tastes, listen to Black Sabbath.

Last.fm's data indicates that Black Sabbath is similar to *Alice Cooper, Deep Purple, Ozzy Osbourne* that are in your profile.

Your friends like Black Sabbath.

Black Sabbath has similar tags as: *Dio, AC/DC* that are in your profile.

Black Sabbath is very popular in the last.fm database with 2.36 million listeners and 94.6 million playcounts.

Black Sabbath is tagged with *rock, seen_live* that are in your profile.

Figure 1: Example of the ranking question (Task 3 of the study) for the recommended artist “Black Sabbath”.

columns/pathways. Based on this finding, we showed the participant Venn diagrams and two very simple forms of pathways among columns, i.e., two cluster dendrograms, one static and one interactive. Figure 2 illustrates an example of the different formats shown to the same participant for the recommended artist “Deep Purple”. The figure included a static cluster dendrogram, which presented all the visualization information at once. The interactive cluster dendrogram initially hid information but allowed participants to interact with the visualization by clicking on blue bullets in the diagram to reveal additional information about the explanation style. Since Venn diagrams can accommodate three different styles, we restricted all the other explanations to show only three styles. To select which three out of the seven offered styles to show, we chose the three styles reported to improve performance in prior work, i.e., user-based, item-based CF, and popularity-based. As before, we randomized the order that we show the different formats. With this task we tested the following hypothesis:

- H_5 : Explanation format significantly correlates with reception (UXP).

In the middle of the study, we also asked the satisficing question: “Please answer “somewhat not persuasive” to this question”. This question, which is the same for all participants, allowed us to remove participants who were not paying attention to the study.

7 RESULTS

We grouped participants in two basic categories: experienced participants who used last.fm prior to the study and novice participants

Perceived Accuracy ($\alpha = 0.96$)	R^2	<i>Est.</i>
The recommended artist represents my tastes.	0.86	1.05
This is an accurate recommendation	0.88	1.05
I like the recommended artist.	0.93	1.06
Perceived Novelty ($\alpha = 0.94$)	R^2	<i>Est.</i>
I have never listened to this artist before.	0.91	1.44
I am aware of the recommended artist.	0.74	1.19
The recommended artist is new to me.	0.91	1.45
Reception (UXP) ($\alpha = 0.93$)	R^2	<i>Est.</i>
(Confidence): This explanation makes me confident that I will like this artist.	0.73	1.04
(Transparency): This explanation makes the recommendation process clear to me.	0.71	1.06
(Satisfaction): I would enjoy using a recommendation system if it presented recommendations in this way.	0.79	1.17
(Persuasiveness): This explanation for the recommendation is convincing.	0.88	1.19

Table 3: Questions for the main study asked to the participants. Again, factors (perceived accuracy, perceived novelty, and UXP) are determined by participant responses to subjective questions. As before, we report R^2 , *Est.*, and Cronbach’s alpha.

that created a last.fm account to participate in the study. As discussed, for each participant, we ran the HyPER model and picked the top three artists that scored the highest. Our framework supports the creation of up to 7 different explanation styles. In our experiments, the first, second, and third artists that we showed to each user was accompanied with 6.2, 6.3, 5.6 different explanation styles on average (values for standard deviations were: 0.67, 0.62, and 0.99 respectively). Next, we report the factors created from the subjective questions along with statistics related to the fit. Then, we report the results of the study and hypothesis testing. Finally, we report differences in the behavior of the experienced and novice users. Significance levels in this section are reported as follows: *** = $p < .001$, ** = $p < .01$, and * = $p < .05$.

7.1 Participants

We collected 212 samples of within-subjects participant data using AMT. Overall, 92% of participants were between 18 and 50 years of age, and 60% were male. Each participant was rewarded with US \$3 as incentive. Satisficing, the practice of gaming research studies, is a legitimate concern for any crowd-sourced platform [18]. We checked the data for satisficing participants by carefully examining input/timing patterns and checking the answer to the satisficing question. Specifically, we checked for repeated responses (e.g., “4” over and over), providing more than two conflicting responses on the items for each factor (e.g., the participant claims an artist is new to her but then also claims she has heard the artist before). After filtering out participants that exhibited satisficing behavior (14 total), there were $N = 198$ samples for analysis. Out of those users, 91 (46%) were considered “experienced” users who already had a last.fm account, where 107 (54%) were “novices” who created a new last.fm account to participate to the study. At the end of this

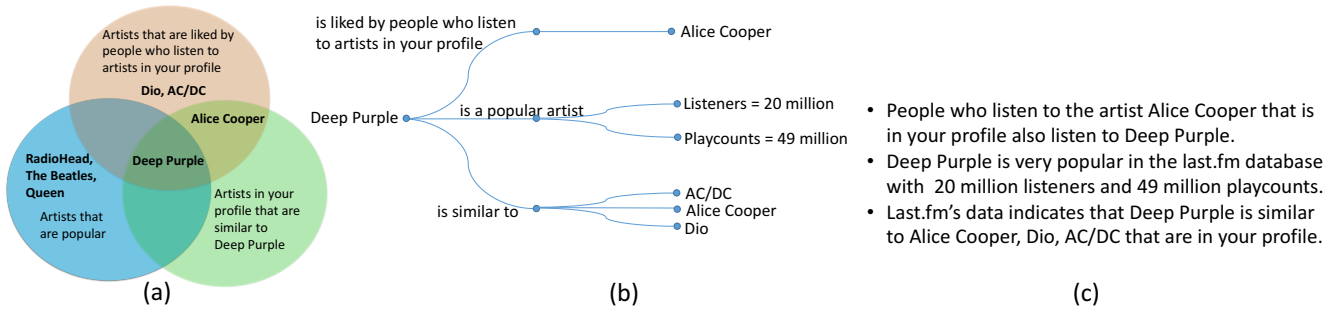


Figure 2: Example of the different explanation formats for the same recommended artist “Deep Purple” (Task 4 of the study). (a) Venn diagrams, (b) static cluster dendrograms, (c) textual. We also showed interactive cluster dendrograms which are the same as static (b) with the difference that the participant can interact with the blue bullets (open or close them).

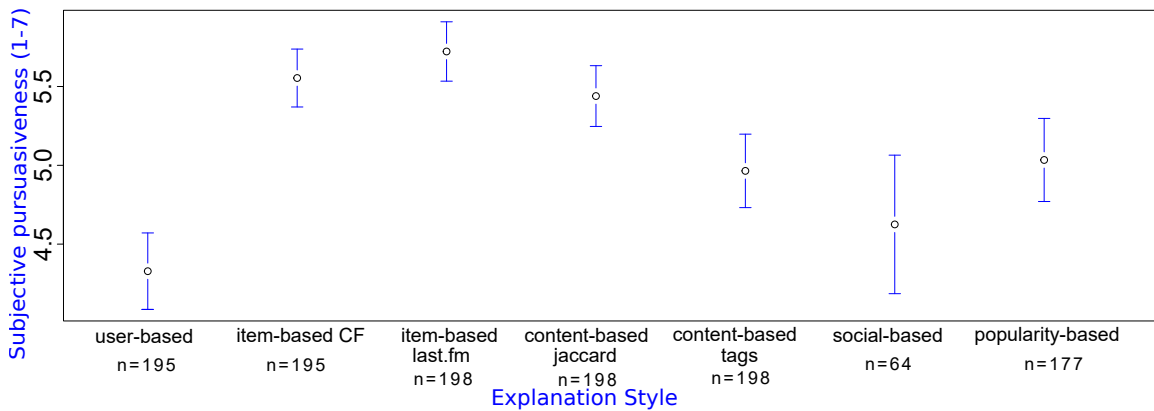


Figure 3: Mean subjective persuasiveness for each style of explanation, taken on a Likert Scale (1-7).

section, we present our findings for these two different groups of participants.

7.2 Factor fit and effectiveness of hybrid recommendations

In Tables 2 and 3 we report the factors that were confirmed from participant responses on the subjective questions (in bold). Next to each factor, we show a measurement of internal reliability (Cronbach’s α [34]) for each dependent variable that was solicited via the questionnaires. All factors achieved good or excellent internal reliability and all factors achieved good discriminant validity using the Campbell & Fiske test [8]. To improve modeling of personality traits (which were not factored), we load a different latent variable on each response with a fixed value (1). Then, we free the variance of each response and fix the variance of the latent variable to the variance of the response.

To validate the quality of the recommendations generated by the HyPER framework, we first asked the participants questions related to perceived accuracy and novelty of the recommendation (reported on top of Table 3) for each recommended artist during the associated task. For each of the three hybrid recommendations, we averaged together the subjective accuracy. The hybrid recommendations resulted in a mean accuracy of 5.64 out of 7 and the best

fitting item for perceived accuracy was “I like the recommended artist”. Working similarly, the mean novelty of the recommendations was 2.1 out of 7 and the best fitting item was “I have never listened to this artist before”. In our analysis below, we use the accuracy and novelty as controlling variables and study their effect on explanations. We note that perceived accuracy and novelty significantly co-varied (-0.355^{***}) in the factor model used to assess factor reliability (see Table 3). This agrees with studies showing that users trust recommender systems more when they receive recommendations that they are familiar with [16].

7.3 Preferences for explanation styles

We used the questions asked in Task 1 of the study to test for differences in persuasiveness when showing different explanation styles. Figure 3 shows the mean subjective persuasiveness (“How persuasive is this explanation?”) across each explanation style. A repeated-measures ANOVA showed a general difference between explanation styles ($F = 32.635, p < 0.0001$). Thus, we accept H_1 , i.e., explanation style significantly correlates with perceived persuasiveness. A Tukey post-hoc test [37] showed significant improvements by item-based CF, content-based Jaccard, and item-based last.fm styles over user-based ($\forall p < 0.001$), popularity-based ($\forall p < 0.025$), content-based tags ($\forall p < 0.001$), and social-based ($\forall p < 0.001$).

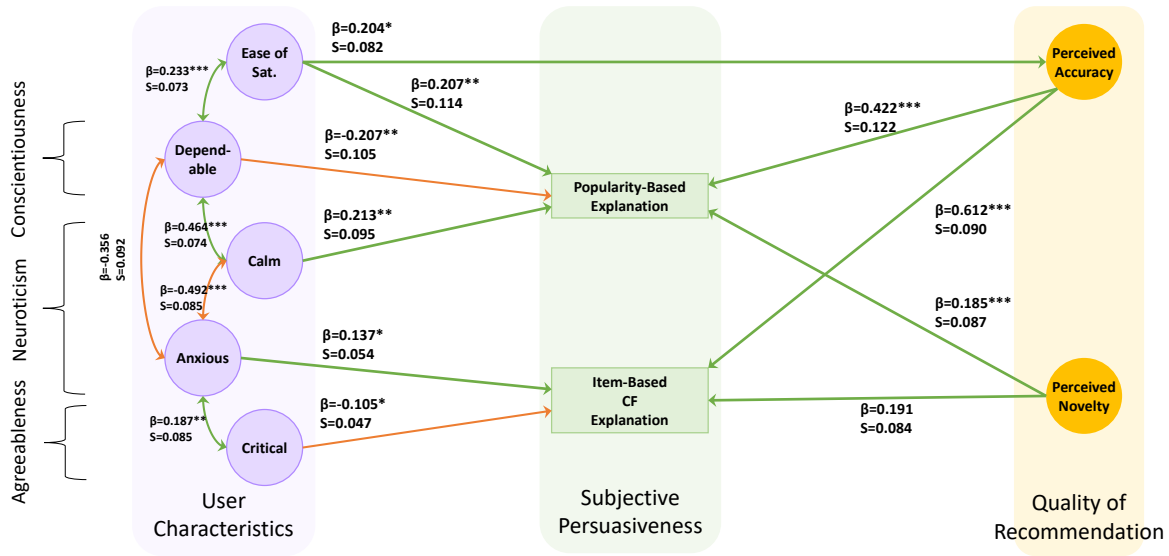


Figure 4: An SEM explaining the role of personality in persuasiveness of explanation. Unidirectional arrows indicate regression, bidirectional arrows indicate covariance; red arrows indicate a negative effect, green arrows indicate a positive effect; latent factors were scaled so β values indicate effect sizes in units of standard deviations. Standard error (S) is given. Model fit: $N = 177$ with 40 free parameters = 4.5 participants per free parameter, $RMSEA = 0.077$ ($CI : [0.060, 0.094]$), $TLI = 0.932$, $CFI > 0.948$ over null baseline model, $\chi^2(80) = 164.628$.

No significant improvement was found for item-based last.fm over item-based CF, or content-based Jaccard.

To test the significance of personality traits in the persuasiveness of an explanation, we conducted an exploratory structural equation modeling (SEM) [38] analysis. It is well known that people may change their ratings of items based on user experience or persuasive explanations [15], so we accounted for this effect by controlling for the accuracy/novelty of each recommendation and the participant’s self-reported ease of satisfaction [32]. Then, we tested for an effect of each of the ten personality traits on the seven different explanation styles by performing a regression between each. This resulted in a total of 70 hypotheses, so we controlled for multiplicity via the Benjamini-Hochberg procedure with $Q = 0.10$ [4], which is recommended for exploratory SEM analysis [10].

Figure 4 shows the results from the exploratory analysis. Of the ten personality traits, only four were shown to have a significant bearing on persuasiveness of the explanation (dependable, calm, anxious, critical). These four responses could be grouped into their larger personality traits: conscientiousness (dependable), neuroticism (anxious, calm), and agreeableness (critical). Conscientious participants reported being easier to satisfy. The participants seemed to be split in terms of neuroticism: calm participants tended to be more receptive of popularity-based explanations while anxious tended to be more receptive of item-based CF explanations. If the participant identified as dependable *and* calm or anxious *and* critical, the effects disappeared. As a result, we accept H_2 , i.e., a person’s personality significantly correlates with perceived persuasiveness. Finally, the effect sizes of perceived accuracy appeared to be double that of perceived novelty and any personality-based effect.

7.4 Preferred number of explanation styles

Next, we analyzed the orderings given by the participants in the ranking questions (Task 2 of the study). First, we noted that if the rankings are treated as ratings (1st position = 7 points, 2nd position = 6 points, etc.), each explanation style has the same relative score as shown in Figure 3 (this serves as a second level of validation for explanation preferences). Second, the mean number of explanation styles ranked was 2.61. However, we found that almost 40% of participants chose to leave all explanation styles in the bottom box without ranking them. After removing these participants, the mean number of explanation styles was 4.32. To test H_3 (people prefer to see the maximum number of explanation styles available), we conducted a one-sample t-test to check if the mean of the sample was significantly different than 7, which was the maximum number of available explanation styles. We found a significant difference ($t = -22.9, p < 0.001$), which remained significant when omitting participants who had not ranked any explanations ($t = -13.8, p < 0.001$). Thus, we reject H_3 , concluding that people lose interest after approximately three to four explanation styles. Finally, we tested whether or not personality significantly correlated with the number of explanation styles ranked. We tested ten regressions (multiplicity control again with $Q = 0.10$) within an SEM which revealed that dependable people were likely to rank less ($\beta = -0.166^*, S = 0.15$) and open people were likely to rank more ($\beta = 0.212^* *, S = 0.144$). Thus we accept H_4 , i.e., a person’s personality significantly correlates with their preferred number of explanation styles.

7.5 Textual vs. visual format

As discussed, in Task 3 of the study, for one artist we showed four different explanation formats (one textual and three visual) and asked

participants to answer a set of UX questions reported in Table 3. We plot the persuasiveness score, which reported the best R^2 value (“This explanation for the recommendation is convincing”), for each visual/textual format in Figure 5. A repeated-measures ANOVA showed a difference between treatments ($F = 10.13, p < 0.001$). Therefore, we accept H_5 , i.e., explanation format significantly correlates with a user’s reception of an explanation. Specifically, text explanations were perceived as more persuasive than every visual format ($\forall p < 0.001$). To investigate further, we considered whether visualization familiarity significantly correlated with better reception of the visual formats. Four regressions were tested in an SEM when controlling for the accuracy of the recommendation and self-reported ease-of-satisfaction, showing that more familiarity with visualization significantly correlated with better reception of the Venn diagram ($\beta = 0.151^*, S = 0.077$). Finally, our analysis does not show any statistically significant difference between the static and interactive version of cluster dendrograms.

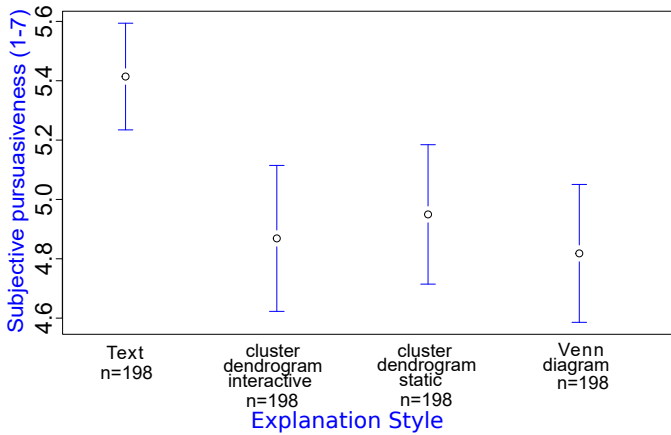


Figure 5: Mean subjective persuasiveness for each format of explanation, taken on a Likert Scale (1-7).

7.6 Experienced vs. novice last.fm users

As discussed 46% of the users participated in the study already had a last.fm account, while 54% of the users created a last.fm account to participate to our study. A regression analysis done in a pathway model showed these two groups of participants varied in three significant ways. In particular, novice users (1) reported much higher perceived accuracy ($B = 0.576 * **$), (2) preferred the content-based Jaccard explanations more ($B = 0.544^*$), (3) and vastly preferred the popularity-based explanations ($B = 0.891 **$). A possible explanation for the lower accuracy of the recommendations for the experienced users might be that these users have already explored the online space of music offerings to a greater extent and as a result they are harder to satisfy when compared to novice users.

8 OFFLINE EVALUATION OF THE HYBRID RECOMMENDER

Our user study focused on evaluating explanations from a hybrid recommender system rather than evaluating the performance of

the recommender system relative to other recommendation techniques. In this section we establish that the proposed HyPER model is not only able to provide explanations, but also produces accurate recommendations compared to other popular recommendation methods. To evaluate the performance of this model, we performed an **offline** evaluation of four different collaborative filtering models in an implicit feedback setting (where only listening activity is available). We compared HyPER’s performance to: (a) an algorithm that ranks the recommended artists based on their popularity, (b) the traditional item-based collaborative filtering algorithm using a cosine similarity metric, and (c) collaborative filtering for implicit feedback datasets (we refer to this method as WRMF) [17].⁴ For all algorithms above we used the implementations provided by the MyMediaLite framework [12].

To evaluate the performance in an offline setting, we first crawled a subset of the last.fm dataset. We randomly sampled user profiles, removing users with fewer than 20 artists in their profile, yielding 2,984 user profiles. For each user, we followed the exact same process described in the first phase of the study (Section 6.1) for crawling data, creating the candidate sets, and computing similarities. The crawling process generated a total of 1,083,705 user-artist interactions from 55,641 users (2,984 users along with their friends and similar users) and 74,728 artists (sparsity=99.97%). The number of friendship connections was 122,320 and the total number of tags for all artists was 123,677.⁵

To create the training and test set, for each of the 2,984 users, we crawled the top 20 artists in their profile. We used 15 artists for training while holding out 5 artists to add to the candidate set (which we generated following the process described in Section 6.1). For each user, each algorithm ranked all the items of their candidate set. We evaluated the performance of the algorithms on recall using four different recall thresholds (5, 10, 25 and 50). To compute the recall metric, we computed the proportion of the 5 ground truth artists from the user’s profile ranked in the top 5 (resp. 10, 25, 50) predictions of each model and average across users.

For the baselines we experimented with various parameters and report the results of the best performing model. For the item-based collaborative filtering we experimented with number of neighbors 10, 20, 50 (we got best results for 20 neighbors). For WRMF we experimented with number of factors 10, 30, 45, 60, 80, 100, 150, number of iterations 15 and 30, and regularization values 0.005, 0.015, 0.03, (we got best result for 60 factors, 15 iterations, and 0.015 regularization).

Table 4 presents the results of HyPER and the baselines. We observe that HyPER outperforms the strongest model, which is the item-based collaborative filtering. As a result, the proposed model not only is able to generate explanations, but it can also provide accurate recommendations.

9 DISCUSSION

In this work, we implemented a personalized hybrid recommender engine that combines multiple sources of information and generates recommendations with a variety of explanation styles in

⁴We also experimented with Bayesian personalized ranking from implicit feedback [30] but results were worse than other techniques.

⁵The data are available here: <https://linqs-data.soe.ucsc.edu/public/kouki-iui19/>.

Model	recall@5	recall@10	recall@25	recall@50
HyPER	0.075	0.121	0.220	0.310
WRMF	0.072	0.114	0.197	0.278
Item-based	0.072	0.118	0.214	0.305
Most popular	0.021	0.037	0.070	0.114

Table 4: Recall@5, 10, 25, 50 for PSL and baselines.

real time. We conducted a crowd-sourced user study where users evaluated the persuasiveness of different explanations generated in real-time and personalized for that user’s tastes that varied in style, number, and format. We evaluated the effects of explanation style, number, and format as well as personality characteristics on user preferences for explanations. Furthermore, we showed that our approach improves the accuracy over existing collaborative filtering baselines while also providing interpretable explanations. The most important findings from our study are as follows.

People prefer item-centric but not user-centric or socio-centric explanations. User-based and social-based explanations were rated as relatively less persuasive by the participants. This was regardless of the age of the participant’s last.fm account. Although the non-personalized popularity-based explanations were rated more favorably, they were still significantly less persuasive than the content-based explanations. Relevant literature [13, 15] showed that users can evaluate content-based explanations precisely and that a content-based interface is highly preferable. Our findings in this paper, reinforce the findings that content-based explanations are a good option when a system can provide only one explanation method. Moreover, we found that calm participants (low neuroticism) preferred popularity-based explanations, while anxious participants (high neuroticism) preferred item-based CF explanations. Additionally, participants that identified as dependable did not have any preference for the popularity-based explanation. Likewise, neurotic participants (who were also likely to be introverted and reserved), showed a slight preference for item-based CF explanations.

People prefer to see at most three to four explanation styles. Our analysis when manipulating the number of explanation styles indicated that a relatively large percentage of users prefer to see no explanation with a recommendation, a possible artifact of our experimental design. We plan to investigate this in more detail in our future work. For the rest of the users, we found that the average number of explanations they preferred is 4.32. We also found that open participants were persuaded by many explanations, while conscientious participants preferred fewer. One possible dynamic that might result in these preferences is that open participants are likely to seek new experiences, while conscientious participants may be turned off by clutter. However, despite the significant effects, due to the correlation between those two traits and relatively low effect sizes, personalization of the number of explanations shown to a user may be unnecessary. A default of three to four explanations would likely be sufficient for most people.

Textual explanations are ideal. Our analysis indicated that text explanations were perceived as more persuasive compared to three different visual formats. When considering visualization familiarity as a control variable, we found that users with more

familiarity in visualization are more receptive to the Venn diagrams. Despite this, our model did not predict that participants familiar with visualization would prefer the Venn diagram over the text explanations. At first, this result may seem to contradict recent work on visual recommender interfaces, but a closer look shows that the results are not directly comparable. In particular, Parra et al. [29] proposed Venn diagrams to implement interactive interfaces and showed that they are more engaging to the users when compared to non-controllable ranked lists. However, their work focused on the user controllability aspect, while in our work the focus was on the persuasiveness of explanations in a more isolated and static way. In another study [22], Venn diagrams did not perform significantly different from a variety of text-based explanations. This study, however, was limited to non-personalized “mock” explanations and interfaces, so users were not receiving real recommendations during the task. In the study presented in this paper, users were exposed to real personalized recommendations with explanations so they may have been more focused on evaluating the recommendations with explanations and thus may have seen the visual explanations as an unnecessary hindrance to their assessments. In summary, we believe that in static recommendation contexts, textual explanations would likely satisfy nearly every consumer.

10 LIMITATIONS AND FUTURE WORK

The results presented in this paper are based on a personalized user study in AMT with users and data from the music domain. Although our observations may hold in other domains and contexts as well, we plan to conduct additional studies in order to generalize our results and account for differences in other domains. For example, we found that, in the music domain, people prefer to see three to four explanations. However, in domains such as job recommendations, the decision that a person would make based on a recommendation is of “higher risk” in the sense that it would affect their career instead of what track to listen next. In such cases, we believe that people may prefer a larger number of explanations in order to better understand the reasoning of the recommendations before making a decision.

In our work, we assume that users operate in a context where they have sufficient time and space to explore the recommendations. We plan to further study whether our findings can generalize for the cases when the users operate in time and/or space constrained contexts (e.g., on a mobile device while on the move). We also showed that our hybrid recommender system outperforms other popular ranking methods in an offline setting. In the future, we plan to additionally compare the performance of the proposed framework and the baselines in an online setting. Additionally, in our offline evaluation we compared only to collaborative filtering models, while in our future work we plan to compare to other hybrid models as well [19]. Additionally, we are interested in studying if and under which circumstances it is beneficial that explanations participate in the process of ranking the recommendations. Finally, we plan to support a conversational recommender system that will use the observations from interactions with the user and adjust the explanations based on the user’s preferences.

ACKNOWLEDGEMENTS

This work was partially supported by the National Science Foundation grants CCF-1740850 and IIS-1703331 and by the Defense Advanced Research Projects Agency under contract number FA8650-17-C-7715. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation or the Defense Advanced Research Projects Agency.

REFERENCES

- [1] G. Adomavicius, N. Manouselis, and Y. Kwon. 2015. *Multi-Criteria Recommender Systems*. Recommender Systems Handbook, Second Edition, Springer US.
- [2] I. Andjelkovic, D. Parra, and J. O'Donovan. 2019. Moodplay: interactive music recommendation based on artists' mood similarity. *International Journal of Human-Computer Studies*, 121.
- [3] S. Bach, M. Broecheler, B. Huang, and L. Getoor. 2017. Hinge-loss markov random fields and probabilistic soft logic. *Journal of Machine Learning Research (JMLR '17)* 18, 109.
- [4] Y. Benjamini and Y. Hochberg. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society (JRSS '95)* 18, 109.
- [5] S. Berkovsky, R. Taib, and D. Conway. 2017. How to recommend?: user trust factors in movie recommender systems. In *Proceedings of the 22nd International Conference on Intelligent User Interfaces (IUI '17)*.
- [6] M. Bilgic and R. Mooney. 2005. Explaining recommendations: satisfaction vs. promotion. In *Beyond Personalization Workshop in conjunction with International Conference on Intelligent User Interfaces (IUI '05)*.
- [7] S. Bostandjiev, J. O'Donovan, and T. Höllerer. 2012. Tasteweights: a visual interactive hybrid recommender system. In *Proceedings of the 6th ACM Conference on Recommender Systems (RecSys '12)*.
- [8] D. Campbell and D. Fiske. 1959. Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, 56, 2.
- [9] S. Chang, F. Harper, and L. Terveen. 2016. Crowd-based personalized natural language explanations for recommendations. In *Proceedings of the 10th ACM Conference on Recommender Systems (RecSys '16)*.
- [10] R. Cribbie. 2007. Multiplicity control in structural equation modeling. *Structural Equation Modeling*, 14, 1.
- [11] G. Friedrich and M. Zanker. 2017. A taxonomy for generating explanations in recommender systems. *AI Magazine*, 32, 3.
- [12] Z. Gantner, S. Rendle, C. Freudenthaler, and L. Schmidt-Thieme. 2011. My-medialite: a free recommender system library. In *Proceedings of the 5th ACM Conference on Recommender Systems (RecSys '11)*.
- [13] F. Gedikli, D. Jannach, and M. Ge. 2014. How should i explain? a comparison of different explanation types for recommender systems. *International Journal of Human-Computer Studies*, 72, 4.
- [14] S. Gosling, P. Rentfrow, and W. Swann. 2003. A very brief measure of the big-five personality domains. *Journal of Research in Personality*, 37, 6.
- [15] J. Herlocker, J. Konstan, and J. Riedl. 2000. Explaining collaborative filtering recommendations. In *Proceedings of the Conference on Computer Supported Cooperative Work (CSCW '00)*.
- [16] J. Herlocker, J. Konstan, L. Terveen, and J. Riedl. 2004. Evaluating collaborative filtering recommender systems. *ACM Transactions on Information Systems (TOIS '04)* 22, 1.
- [17] Y. Hu, Y. Koren, and C. Volinsky. 2008. Collaborative filtering for implicit feedback datasets. In *IEEE International Conference on Data Mining (ICDM '08)*.
- [18] P. Ipeirotis. 2010. Mechanical turk: now with 40.92% spam. <http://www.behind-the-enemy-lines.com/2010/12/mechanical-turk-now-with-4092-spam.html>. Blog. (2010).
- [19] D. Jannach, I. Kamehkhosh, and L. Lerche. 2017. Leveraging multi-dimensional user models for personalized next-track recommendation. In *Proceedings of the Symposium on Applied Computing (SAC '17)*.
- [20] B. Knijnenburg, S. Bostandjiev, J. O'Donovan, and A. Kobsa. 2012. Inspectability and control in social recommenders. In *Proceedings of the 6th ACM Conference on Recommender Systems (RecSys '12)*.
- [21] P. Kouki, S. Fakhraei, J. Foulds, M. Eirinaki, and L. Getoor. 2015. Hyper: a flexible and extensible probabilistic framework for hybrid recommender systems. In *Proceedings of the 9th ACM Conference on Recommender Systems (RecSys '15)*.
- [22] P. Kouki, J. Schaffer, J. Pujara, J. O'Donovan, and L. Getoor. 2017. User preferences for hybrid explanations. In *Proceedings of the 11th ACM Conference on Recommender Systems (RecSys '17)*.
- [23] T. Nguyen, D. Kluver, T. Wang, P. Hui, M. Ekstrand, M. Willemsen, and J. Riedl. 2013. Rating support interfaces to improve user experience and recommender accuracy. In *Proceedings of the 7th ACM Conference on Recommender Systems (RecSys '13)*.
- [24] X. Ning, C. Desrosiers, and G. Karypis. 2015. *A comprehensive survey of neighborhood based recommendation methods*. Recommender Systems Handbook, Second Edition, Springer US.
- [25] I. Nunes and D. Jannach. 2017. A systematic review and taxonomy of explanations in decision support and recommender systems. *User Modeling and User-Adapted Interaction (UMUAI'17)* 27, 3-5.
- [26] J. O'Donovan, B. Smyth, B. Gretarsson, S. Bostandjiev, and T. Höllerer. 2008. Peerchooser: visual interactive recommendation. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '08)*.
- [27] S. Oramas, L. Espinosa-Anke, M. Sordo, H. Saggion, and X. Serra. 2016. Information extraction for knowledge base construction in the music domain. *Data & Knowledge Engineering*, 106, C.
- [28] A. Papadimitriou, P. Symeonidis, and Y. Manolopoulos. 2012. A generalized taxonomy of explanations styles for traditional and social recommender systems. *Data Mining and Knowledge Discovery*, 24, 3.
- [29] D. Parra, P. Brusilovsky, and C. Trattner. 2014. See what you want to see: visual user-driven approach for hybrid recommendation. In *Proceedings of the 19th International Conference on Intelligent User Interfaces (IUI '14)*.
- [30] S. Rendle, C. Freudenthaler, Z. Gantner, and L. Schmidt-Thieme. 2009. BPR: bayesian personalized ranking from implicit feedback. In *Proceedings of the 25th Conference on Uncertainty under Artificial Intelligence (UAI '09)*.
- [31] M. Sato, B. Ahsan, K. Nagatani, T. Sonoda, Q. Zhang, and T. Ohkuma. 2018. Explaining recommendations using contexts. In *Proceedings of the 23rd International Conference on Intelligent User Interfaces (IUI '18)*.
- [32] J. Schaffer, J. O'Donovan, and T. Höllerer. 2018. Easy to please: separating user experience from choice satisfaction. In *Proceedings of the 26th Conference on User Modeling, Adaptation and Personalization (UMAP'18)*.
- [33] P. Symeonidis, A. Nanopoulos, and Y. Manolopoulos. 2009. Movieexplain: a recommender system with explanations. In *Proceedings of the 3rd ACM Conference on Recommender Systems (RecSys '09)*.
- [34] M. Tavakol and R. Dennick. 2011. Making sense of cronbach's alpha. *International Journal of Medical Education*, 2.
- [35] N. Tintarev and J. Masthoff. 2012. Evaluating the effectiveness of explanations for recommender systems. *User Modeling and User-Adapted Interaction (UMUAI'12)* 22, 4-5.
- [36] M. Tkalcić and L. Chen. 2015. *Personality and Recommender Systems*. Recommender Systems Handbook, Second Edition, Springer US.
- [37] J. Tukey. 1949. Comparing individual means in the analysis of variance. *Biometrics*.
- [38] J. Ullman and P. Bentler. 2003. *Structural equation modeling*. Wiley Online Library.
- [39] K. Verbert, D. Parra, P. Brusilovsky, and E. Duval. 2013. Visualizing recommendations to support exploration, transparency and controllability. In *Proceedings of the 18th International Conference on Intelligent User Interfaces (IUI '13)*.
- [40] J. Vig, S. Sen, and J. Riedl. 2009. Tagsplanations: explaining recommendations using tags. In *Proceedings of the 14th International Conference on Intelligent User Interfaces (IUI '09)*.