Generating and Understanding Personalized Explanations in Hybrid Recommender Systems

PIGI KOUKI[*], relational AI

JAMES SCHAFFER, Sysco Corporation

JAY PUJARA, University of Southern California

JOHN O'DONOVAN, UC Santa Barbara

LISE GETOOR, UC Santa Cruz

Recommender systems are ubiquitous, and shape the way users access information and make decisions. As these systems become more complex, there is a growing need for transparency and interpretability. In this paper, we study the problem of generating and visualizing personalized explanations for recommender systems which incorporate signals from many different data sources. We use a flexible, extendable probabilistic programming approach, and show how we can generate real-time personalized recommendations. We then turn these personalized recommendations into explanations. We perform an extensive user study to evaluate 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. First, we evaluate the performance of the recommendations in terms of perceived accuracy and novelty. Next, 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 also 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. Finally, we perform a post analysis which shows different preferences for explanation styles between experienced and novice last.fm users.

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

Additional Key Words and Phrases: 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. Generating and Understanding Personalized Explanations in Hybrid Recommender Systems. 1, 1 (March 2019), 39 pages. https://doi.org/10.1145/3301275.3302306

1 INTRODUCTION

Recommender systems are an indispensable tool for consumers to help them navigate the vast number of options for content and products. As recommendations have become central to combating information overload and shaping

[*] Part of this work was performed while this author was a graduate student at UC Santa Cruz.

Authors' addresses: Pigi Kouki, relational AI, pigi.kouki@relational.ai; James Schaffer, Sysco Corporation, schaffer.james@corp.sysco.com; 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.

decisions, users increasingly demand convincing explanations that will help them understand why they were presented with a given recommendation [7, 32]. The increasing complexity of recommender systems has fueled a corresponding need for explanations to evolve in order 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., [18, 6, 40, 52, 11]) 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. Each hybrid recommendation is the result of the fusion of many knowledge sources, and choosing which sources to expose to a user presents a new research challenge. A recent survey [35] of different single-style explanation approaches 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 generated for hybrid systems.

In this paper, we present a hybrid recommender system that can incorporate a variety of recommendation approaches (e.g., content-based and CF). Using this hybrid system, we generate 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 conduct a large personalized user study on a real system from the last.fm music domain. We experimentally vary 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, in certain cases, there is a relationship between a person's personality and the type of explanation that is most persuasive. To understand the effects of personality traits, we also conduct an analysis for personalized hybrid explanations. To the best of our knowledge, our work is the first study on the effect of such variables on *personalized hybrid explanations*.

In this work, we extend our earlier hybrid recommender system [23] to produce real-time recommendations while incorporating a variety of information sources. Moreover, we present 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 use these signals to create a hybrid model that incorporates user-user and item-item similarities using CF, content, social, and popularity information. We also showcase a translation system that generates customized explanations from the output of the hybrid system in real time. More specifically, we convert 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. 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 generate real-time recommendations along with personalized explanations for users of the last.fm music platform. We conduct a crowd-sourced user study ($N = 198$) using Amazon Mechanical Turk (AMT), recruiting users with active last.fm accounts. Our first goal is 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 study the effect of personality traits of the users. We find

Explanation Style	We recommend *U2* because:
(I) User-based	User *Aren* with whom you share similar tastes in artists, listens to U2.
(II) Item-based	(a) People who listen to your profile item *AC/DC* also listen to U2. (b) Last.fm's data indicates that U2 is similar to *Coldplay* that is in your profile.
(III) Content	(a) U2 has similar tags as *Beatles* that is in your profile. (b) U2 is tagged with *rock* that is in your profile.
(IV) Social	Your friend *Cindy* likes U2.
(V) Item popularity	U2 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 (U2). Multiple styles are generated from the hybrid model (the first four are personalized, while the fifth one is non-personalized).

interesting patterns between explanation persuasiveness of particular styles and personality traits which we analyze in our results. Second, we study 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 view explanations in visual or textual formats?

Additionally, during our study, we evaluate the recommendations provided by our method in terms of perceived accuracy and novelty. In order to evaluate the effectiveness of our approach, we compare the accuracy and novelty of the proposed framework to a random recommendation baseline. We find that our recommendations are 59.8% more likely to be perceived accurate compared to the baseline. Finally, during our post hoc analysis we find that we can divide participants into two subgroups: (a) participants that were already using the last.fm music recommendation platform (i.e., experienced last.fm users) and (b) participants that created a last.fm account for the sake of participating to our study (i.e., novice last.fm users). For each of the questions defined above, we report results for these two different subgroups. We find that the two subgroups differ in terms of perceived accuracy and novelty and at the same time they report different preferences for explanation styles.

We note that this work is an extended version of Kouki et al. [24]. Our contributions include: (1) a hybrid recommender system that can provide real-time recommendations with up to seven different explanations styles in various formats (such as textual and visual), (2) an online evaluation of the proposed system in terms of perceived accuracy and novelty of the recommendations, (3) insights regarding the most persuasive styles, the ideal number of explanation styles, and the most persuasive presentation formats, and (4) insights regarding the effect of different personality traits in persuasiveness of explanations, and (5) insights into how the preferences of experienced vs. novice users vary in terms of recommendation styles. Contributions 2, 3, 4, and 5 are the result of a personalized user study performed on real users of the last.fm music platform.

The remainder of the paper is structured as follows. In Section 2 we discuss the related work. In Section 3 we describe our hybrid recommendation framework. In Section 4 we describe the properties of the last.fm dataset. In Section 5 we define the research questions we aim to answer. In Section 6 we give an overview of the user study and in Section 7 we present the results which answer the research questions. In Section 8 we summarize our findings. Finally, in Section 9 we discuss the limitations of this work and our future research plans.

2 RELATED WORK

There is an emerging body of work on explanations for recommender 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. We organize our discussion along three themes: i) work on single-style explanations, ii) work combining more than one explanation style and, iii) work surveying the state-of-the-art in explanations for decision support and recommender systems.

Single-style explanations. The basic difference between our work and the papers below is that we focus on hybrid explanations while all these works provide single-style explanations. For the cases that the related works is very similar to ours, we provide additional details about the differences. Herlocker et al. [18] showed that certain explanation and presentation styles can increase the persuasiveness of a recommender system in a user-based CF setting, i.e., a recommender system's effectiveness in convincing users to make a purchase. They also showed that users value the explanations and would like to see them in typical recommender systems. Bilgic and Mooney [6] compared single-style explanations that use content-based keywords, item-based CF, or prior rating history. They showed that the first approach decreases effectiveness while the last two approaches increase effectiveness. Vig et al. [52] introduced tagsplanations which provide explanations using community tags. The authors studied two aspects of tag-based explanations: the relationship of the tag to the recommended item and the relationship of the user to the tag. The experimental evaluation showed that both tag relevance and tag preference help improve justification, effectiveness, and mood compatibility. Chang et al. [11] admitted that explanations coming directly from algorithms might be very simplistic and possibly not convincing. They proposed to leverage the wisdom of the crowd to generate personalized natural language explanations. Comparison of textual explanations to tag-based explanations showed that the former is superior in increasing the trust and satisfaction of the users to the system. Muhammad et al. [30] considered that explanations should play a significant role in ranking the recommendations, so they proposed a solution that generates features from reviews that refer to a positive or negative meaning and use these to rank the list of the recommendation results. Tintarev and Masthoff [43] studied the effect of explanations in effectiveness and satisfaction, as well as the trade-off between the two. The authors found that, despite improving user satisfaction, personalization can reduce the effectiveness of content-based explanations. Gedikli et al. [16] 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. Lu et al. [27] proposed a new recommendation model that extracts information about topics from reviews. To achieve this, the authors implemented a recurrent neural network method with an attention mechanism. The generated textual features from the user-generated content were used to explain recommendations.

Recently, Oramas et al. [34] proposed a new method to build a music knowledge base entirely from scratch and they extracted relationships between pairs of entities by combining a state-of-the-art linking tool with a rule-based algorithm. The result of this process revealed a number of facts that were not known in common knowledge repositories (e.g., Wikipedia). The music knowledge-base along with the new facts were used to generate content-based natural language explanations. One of the main focuses of this work was to evaluate how good the proposed system was in explaining relationships between songs. As a result, the authors performed a user study to directly ask participants about explanations on song recommendations. To rank the songs, the paper used a simple baseline approach that recommends songs similar to the ones a user likes; the explanations provided were content-based only. A small user study that involved 35 participants showed that for musically untrained users, providing explanations improves the user experience. Our approach does not involve knowledge graphs. Instead, the focus is on generating hybrid explanations from a state-of-the-art recommender system and understanding the user preferences in various settings (i.e., style, ranking, visualization).

The literature has also proposed several interfaces that support single style explanations. PeerChooser [33] 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. Specifically, the authors analyzed the differences across users with respect to the Big Five personality traits. This work studied the ways the items are recommended, thus it did not involve an algorithm to generate recommendations. Inspired by this work, we also use the personality traits; in our case, we study whether varying the explanation style or number of explanations changes the persuasiveness of explanations when controlling for different personality traits.

Recently, Millecamp et al. [28] performed an online user study using the Spotify music platform. The work focused on studying two different versions of a simple recommender system: one version that did not support explanations and one version that supported visual explanations using bar charts and scatterplots. The paper studied the effect of five specific personal characteristics (musical sophistication, visual working memory, locus of control, need for cognition, and visualisation literacy) on user perception and interaction with the system. The evaluation results showed that (1) users with a low need for cognition are more confident when recommendations were explained, the opposite is true for users with high need for cognition, (2) participants play more songs per minute in the version without explanations, (3) users with a higher musical sophistication or lower visualisation literacy tend to listen to fewer songs when evaluating them, (4) a lower visualisation literacy results in a higher precision. Finally, the paper concluded that (a) explanations should be personalized, (b) the users should select whether to be provided with explanations, and (c) the system should be able to show varying levels of explanation detail based on users' preferences. The most important differences between this work and ours are the following: the work of Millecamp et al. [28] involved a simple recommendation algorithm, where the user selected a specific artist and some attribute values as seeds and then recommendations were queried from Spotify. In our case, the recommendations are generated by a state-of-the-art recommendation engine that takes into account a variety of signals from a user's profile (e.g., songs, social connections). Another difference is that, in our case, we study the effect of hybrid explanations while this work studied explanations of two specific visualizations (bar charts and scatterplots) that are more comparable to the tag-based explanations (one out of the seven different styles that we provide in our case). Finally, a major difference is the fact that the user personality characteristics studied are different.

Hybrid explanations. Recent approaches studied the effect of hybrid explanations and how to best present those explanations that involve more than one explanation style, i.e., the focus is on proposing graphical interfaces to visualize the different explanation styles. TalkExplorer [51] combined content, tag, and social-based filtering techniques to provide an interactive interface in the form of clustermaps. The interface supported both explanations as well as exploration and control of the recommendations. User studies were performed when recommending talks at two conferences. The evaluation results showed that users find that the visualization helped them understand why an item is recommended to them and at the same time it gave them more insight compared to a ranked list of recommendations. The proposed framework offered two distinct recommendation methods: content-based and tag-based. In contrast, our recommendation engine is a hybrid model that includes content-based, tag-based, item-based, user-based, social-based, and popularity-based features. Another major difference is the fact that TalkExplorer focused on offering an interactive visualization tool that allowed users for exploration and control, while, in our case, we focus on understanding the effect of different personalized explanation styles. Moreover, the user study conducted for TalkExplorer involved highly technical end-users (attendees of computer science conferences) which may imply that users were comfortable in using recommender systems while in our case users are only required to use the last.fm music platform. Also, the recommending items were research talks and papers, while, in our case, the domain is music recommendation. Finally, TalkExplorer did not take into account personal characteristics of the users.

SetFusion [36] built on TalkExplorer and replaced clustermaps with Venn diagrams showing improved user experience. The evaluation was conducted again in the research talk recommendation setting. Another important finding was that a controllable interface is more engaging for users compared to a non-controllable one when having conference attendees as users. In more detail, SetFusion offered a visual interface for hybrid recommender system that could be controlled. Users could manually fuse and control the importance of different recommendation methods and then inspect the fusion results using an interactive visualization based on Venn diagrams and sliders. There were three different recommendation methods supported: content-based, author-based popularity, and bookmarking popularity. The final recommendation was a weighted average of the above methods while at the same time, the user could change the weight of each method based on his/her personal interests. There are three important differences between SetFusion and our method: (1) Setfusion focused on the user controllability during the process of recommending new items. In our work, we do not study this aspect of recommender systems; we rather focus on understanding user preferences on different explanation styles, their number, and format. (2) SetFusion involved three specific recommendation algorithms and as a result three explanation styles, while our algorithm involves seven different recommendation methods and explanation styles. Also in our case, the recommender system is not a simple weighted average of the methods but a complex graphical model based on a powerful probabilistic programming language. (3) Similar to the case of TalkExplorer, users in SetFusion were highly technical since they were attendees of computer science conferences that received research talk and paper recommendations while, in our case, users use last.fm platform and received music recommendations.

IntersectionExplorer [10] proposed a new tool that allowed for users to explore the provided recommendations in a multi-perspective way (personal, social, and content relevance). The proposed tool mixed recommendations from four different recommender systems (tag-based, bookmark-based, external bookmark-based, bibliography-based) along with data generated by humans (e.g., tags, bookmarks) and provided the users with a visualization that allowed for exploration. The tool was implemented using UpSet [26], a scalable visualization technique for analysis of sets, intersections, and aggregates of intersections. The framework was evaluated in the domain of conference paper recommendations and the results showed that exploration helps users find relevant papers. IntersectionExplorer shared the same core concepts with TalkExplorer and SetFusion, i.e., the focus was on exploring multiple perspectives of relevance. The basic difference is that IntersectionExplorer was built using UpSet, while TalkExplorer used a cluster map visualization and SetFusion used Venn diagrams. A basic difference between Intersection Explorer and our approach is that in Intersection Explorer the recommendations coming from users with similar tastes were not part of a hybrid recommendation engine. Additionally, our approach differs with IntersectionExplorer in the same way that it differs with TalkExplorer and SetFusion described above.

TasteWeights [7] built an interactive hybrid recommender system that combined social, content, and expert information. The framework showed the reasoning behind the recommendations in the form of pathways among columns. In addition to TasteWeights, our approach uses signals that come from pure CF information (both user and item-based) to produce recommendations. CF algorithms have been proved to show high accuracy when the user-item recommendation matrix is sparse, while content and social-based filtering have been shown to work well in cold-start settings. The combination of both has been shown to provide better accuracy compared to the individual approaches. Another difference is that in TasteWeights predictions were made independently from each source before they were combined using a hybrid strategy. In our approach, we define one single model which incorporates all different information sources so there is no need to perform any hybrid strategy.

Tsai and Brusilovsky [46] implemented a visual interface called Relevance Tuner that presented recommendations beyond a ranked list and at the same time focused on diversity. The result of the user study in social recommendation at academic conferences showed that visual interfaces help users complete specific tasks with reduced exploration, while it also allows them to perceive the recommendation diversity. Next, Tsai and Brusilovsky[47] extended Relevance Tuner by adding explanations. The new system called Relevance Tuner+ was an interactive hybrid social recommender system that incorporated five recommender methods where the weight of each method in the final recommendation could be adjusted by the user. The system provided explanations in the form of six different interfaces: Venn word clouds, topic similarity, co-authorship graph, interest similarity, geographic distance, and relevance equation. Each of the five first interfaces provided a single explanation coming from a single recommender component. Only the relevance equation provided a hybrid explanation which was an equation about the percent of each of the five recommender components. The system was evaluated in the context of social recommendation at academic conferences. The results showed that users leverage the sliders to adjust source weights which indicates that an interactive interface improves user experience and initiates user-driven exploration. At the same time, the study showed that explanations are not used as heavily. Among the six different visual explanations, the Venn word clouds and the co-authorship graph ranked the highest in terms of understandability, persuasiveness, and enjoyability. One additional finding of the study was that the perception of system explainability improves when users can inspect the social recommendations with an on-demand explanation but, at the same time, the user perception of control and the sense of ease of use is reduced. The major differences between this work and ours are the following: In Relevance Tuner+ five of the six visualization interfaces were of single style, while in our work, all visualizations provide explanations of multiple styles (the sixth visualization simply provided an equation about the weight of each method to the final recommendation). Another major difference is that the goal in Relevance Tuner+ was to provide explanations in an interactive interface. In our work, the recommendations are static and the users cannot interact with the system, since the focus is to study the effect of different explanation styles, number, and format. Studying our system in an interactive environment is left to our future work. Additionally, Relevance Tuner+ was evaluated in a social recommendation setting using a small number of users (N=33) which are conference attendees while in our case the study is on music recommendations using a larger number of users (N=198) that are using the last.fm music platform. Finally, another difference is that the recommendations from Relevance Tuner+ were computed as a weighed average of different methods while, in our case, the recommendations are computed from a graphical model.

Symeonidis et al. [40] combined content-based filtering and rating history to generate natural language explanations which were all of the same type in the movies domain. In more detail, the authors implemented a framework called MoviExplain that used a simple heuristic to interpret a rating by a user to a movie, as a vote to the features of the movie (such as actors or directors). Based on these features, MoviExplain built a feature profile for each user and then grouped similar users in the same clusters. The generated features reflected the tastes of a group as a whole. A basic difference between MoviExplain and our proposed framework is that, in our case, the hybrid recommendation engine combines a variety of information sources (including social, content, tags, collaborative-filtering) while in the case of MoviExplain the recommendation engine used only user rating profiles along with item profiles. Another difference is the following: the explanations provided by MoviExplain had always the same form: *"Movie X is recommended because it contains features a, b, ... which are also included in movies Z, W, ... you have already rated"*. Contrary to that, our method can provide up to seven different explanation styles.

Sato et al. [37] proposed to use contexts for explanations (e.g., usage scenarios or accompanying persons). The proposed recommender system needed to select the most relevant pairs of contexts and items for the users. To achieve this the system needed to compute three different matches: (a) a user-item match, (b) an item-context match, and (c) a user-context match. The recommender system was implemented using field factorization machines [21]. The proposed framework could provide seven different explanations: four single style explanations (baseline, demographic, content, context) and three hybrid explanations (demographic and content, content and context, demographic and content and context). To evaluate the system the authors conducted a user study in the restaurant recommendation domain where users were asked to evaluate the explanations on their persuasiveness and usefulness. The results of the study showed that context style explanations are better compared to demographic. In addition, the study showed that combining context with other styles of explanations (e.g., demographic) further improves persuasiveness and usefulness. This work focused on context-based explanations and combination of context style with other styles. In our work, we do not take context into account - we leave this to our future work. Another difference is that the system of Sato et al. [37] compared seven different explanation styles, but these explanation styles could not be combined together to form one hybrid explanation as in our case. The highest number of hybrid explanations that could be provided was three, while in our case the total number is seven. Finally, the authors compared only textual explanations while we also consider three different visualizations.

Recently, Andjelkovic et al. [2] implemented MoodPlay, a hybrid music recommender system that combined content and mood-based filtering. The system provided interaction, control, and explanations. Explanations were provided using links to the last.fm music profile of the recommended items and using circles representing mood. The hybrid music recommender system incorporated information about mood and content. The system could rerank the recommendations based on a user's mood. One of the basic findings of the user study in the music domain was that visualization of mood information in a visual space significantly improves users' understanding of recommendations. This work focused on providing an interactive visual interface that could produce recommendations that take mood into account. Contrary to that, we do not take mood into account. Instead, we focus on providing a hybrid music recommender system that can take into account a variety of information sources, as well as understand user preferences on different explanation styles.

In our previous work [25] 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.

Recently, Millecamp et al. [29] performed an online user study showing explanations of recommendations in two domains: music and cameras. The explanation styles were content, collaborative, or a mix of those. The paper did not find any correlation between the overall user's perception of the system and the different explanations. However, the findings showed that the difference of domains (music vs. cameras) and the difference in need for cognition for the users affect the users' perception of the system. In the music domain, the paper found that (1) users prefer an interface that supports explanations which agrees with our findings and (2) the user's perception of the system is moderated by the user's need for cognition. The aforementioned work showed the users a static list of non-personalized recommendations (in other words there was no recommendation system involved). Contrary to that, in our work the recommendations are a result of a sophisticated hybrid recommender systems and the recommendations are personalized to each user's tastes. The design of the two studies is also different, the study from Millecamp et al. [29] was between-subjects (i.e, different users were shown different explanation styles) while our study is within subjects (i.e., all users were exposed to all the different scenarios).

Finally, Tsukuda and Goto [48] extended the work presented here to diversify both recommended items and explanations. The authors used the framework we propose here to generate hybrid recommendations and explanations. Then, they diversified the recommended items using a state-of-the-art method [13]. Afterwards, the paper introduced a new greedy algorithm to diversify different explanation styles. The basic idea behind the algorithm is that similar sets of explanation styles should not be displayed for similar artists. Experimental evaluation showed that it is possible to diversify both items and explanation styles without largely reducing accuracy.

Surveys on Explanations. Nunes and Jannach [32] 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 [15] proposed a taxonomy that categorizes different explainable recommender systems. The authors argued 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, in this work 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, 29], or uses a baseline recommender [7, 34, 16, 34]. 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 that has been shown to outperform state-of-the-art recommender systems [23]. 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. The hybrid recommender system is implemented using Probabilistic Soft Logic (PSL) [3], a general probabilistic programming language which constructs a probabilistic model using a set of template rules in first-order logic syntax. We choose PSL as our modeling framework for the following two reasons: (1) PSL can incorporate a wide range of signals, such as user-user and item-item similarity measures, content, and social information. (2) PSL models are defined by a set of first-order logical rules and as a result the process of generating and personalizing explanations is transparent; it involves selectively instantiating variable bindings in the rules, and then presenting the personalized, instantiated rules to the user. In what follows, we first describe PSL (Section 3.1). Next, we describe how we use this framework to implement a music recommender system (Section 3.2). Finally, we discuss how we transform the model's probabilistic factors to explanations capturing the different recommender signal types (Section 3.3).

3.1 Hybrid recommendations using Probabilistic Soft Logic

Probabilistic soft logic is a probabilistic programming language that uses a first-order logic syntax to define a graphical model [3]. In contrast to other approaches, PSL uses continuous random variables in the $[0, 1]$ unit interval and specifies factors using convex functions, allowing tractable and efficient inference. PSL defines a Markov random field associated with a conditional probability density function over random variables $\mathbf{Y}$ conditioned on evidence $\mathbf{X}$,

$$P(\mathbf{Y}|\mathbf{X}) \propto \exp\Big(-\sum_{j=1}^{m} w_j \phi_j(\mathbf{Y}, \mathbf{X})\Big) , \tag{1}$$

where ϕ_j is a convex potential function and w_j is an associated weight which determines the importance of ϕ_j in the model. The potential ϕ_j takes the form of a *hinge-loss*:

$$\phi_j(\mathbf{Y}, \mathbf{X}) = (max\{0, \ell_j(\mathbf{X}, \mathbf{Y})\})^{p_j} \ . \tag{2}$$

Here, ℓ_j is a linear function of $\mathbf{X}$ and $\mathbf{Y}$, and $p_j \in \{1, 2\}$ optionally squares the potential, resulting in a *squared-loss*. The resulting probability distribution is log-concave in $\mathbf{Y}$, so we can solve maximum a posteriori (MAP) inference exactly via convex optimization to find the optimal $\mathbf{Y}$. We use the alternating direction method of multipliers (ADMM) approach [8] that has been adapted by Bach et al. [3] to perform this optimization efficiently and in parallel. The convex formulation of PSL is the key to efficient, scalable inference in models with many complex interdependencies. PSL derives the objective function by translating logical rules specifying dependencies between variables and evidence into hinge-loss functions. PSL achieves this translation by using the *Lukasiewicz* norm and co-norm to provide a relaxation of Boolean logical connectives:

$$\begin{aligned} p \wedge q &= \max(0, p + q - 1) \\ p \vee q &= \min(1, p + q) \\ p &= 1 - p \ . \end{aligned}$$

To illustrate how PSL translates logical formula into convex objectives, we provide an example from a music recommendation context. The following rule encodes that users tend to listen to artists of their preferred genres:

$$\textsc{LikesGenre}(u, g) \wedge \textsc{IsGenre}(a, g) \Rightarrow \textsc{Listens}(u, a) \ ,$$

where $\textsc{LikesGenre}(u, g)$ is a binary observed predicate, $\textsc{IsGenre}(a, g)$ is a continuous observed predicate in the interval $[0, 1]$ capturing the affinity of the artist to the genre, and $\textsc{Listens}(u, a)$ is a continuous variable to be inferred, which encodes the probability that the user u likes artist a which is a number between 0 and 1, with higher values corresponding to higher probability that the user will like the artist. For example, we could instantiate u = Jim, g = Rock and a = Metallica. This instantiation results in a hinge-loss potential function in the HL-MRF,

$$\begin{aligned} \max(&\textsc{LikesGenre}(Jim, Rock) \\ &+ \textsc{IsGenre}(Metallica, Rock) \\ &- \textsc{Listens}(Jim, Metallica) - 1, 0) \ . \end{aligned}$$

Recently, Kouki et al. [23] introduced HyPER, a hybrid recommender system that uses PSL. HyPER provides a generic and extensible recommendation framework with the ability to incorporate different recommendation algorithms and consolidate these recommendations using a probabilistic model. To do so, HyPER uses rules and 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. Rules can capture the inputs from different recommender algorithms, item metadata, or user features. 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. In what follows, we present our model in detail.

3.2 Hybrid music recommender model

We propose a hybrid music-recommender system which consists of the below sets of rules where each rule is motivated by a basic principle of a specific recommendation algorithm:[1]

$$\textsc{SimUsers}_{CF}(u_1,\, u_2) \wedge \textsc{Listens}(u_1,\, a) \Rightarrow \textsc{Listens}(u_2,\, a) \tag{3}$$

$$\textsc{SimArtists}_{CF}(a_1,\, a_2) \wedge \textsc{Listens}(u,\, a_1) \Rightarrow \textsc{Listens}(u,\, a_2) \tag{4}$$

$$\textsc{SimArtists}_{last.fm}(a_1,\, a_2) \wedge \textsc{Listens}(u,\, a_1) \Rightarrow \textsc{Listens}(u,\, a_2) \tag{5}$$

$$\textsc{SimArtists}_{content}(a_1,\, a_2) \wedge \textsc{Listens}(u,\, a_1) \Rightarrow \textsc{Listens}(u,\, a_2) \tag{6}$$

$$\textsc{HasTag}(a_1,\, t) \wedge \textsc{HasTag}(a_2,\, t) \wedge \textsc{Listens}(u,\, a_1) \Rightarrow \textsc{Listens}(u,\, a_2) \tag{7}$$

$$\textsc{Friends}(u_1,\, u_2) \wedge \textsc{Listens}(u_1,\, a) \Rightarrow \textsc{Listens}(u_2,\, a) \tag{8}$$

$$\textsc{PopularArtist}(a) \Rightarrow \textsc{Listens}(u,\, a) \tag{9}$$

$$\neg\textsc{Listens}(u,\, a) \tag{10}$$

As an overview, Rule 3 encodes user-based collaborative filtering, Rule 4 is based on an item-based CF signal, Rule 5 is based on a proprietary, domain-specific item similarity algorithm, Rule 6 uses a content-based item similarity, Rule 7 uses a tag-based similiarity, Rule 8 uses a social-based recommendation, Rule 9 is based on a global popularity, and Rule 10 encodes a prior. Although each of these rules is based on a different recommendation approach and can be explained using a different explanation style, the HyPER model integrates these signals into a single prediction for each user, item pair. We now explain each rule in greater detail.

Rule 3 is motivated by the basic principles of the user-based collaborative filtering approach and captures the intuition that similar users like similar artists. An atom such as $\textsc{Listens}(u_2, a)$ 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. In other words, this rule will cause the PSL model to assign a low probability to states where similar users do not listen to similar artists. Mathematically, this reduces the probability of the predicted state as the difference between $\textsc{Listens}(u_2, a)$ and $\textsc{Listens}(u_1, a)$ increases for similar users u_1 and u_2. Atom $\textsc{SimUsers}_{CF}(u_1, u_2)$ 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). However, similarities in this model can be calculated with many different similarity measures. To compute CF similarities in our model, we use Jaccard and cosine similarities over artists in the user's listening history. Jaccard similarity between two users is computed using the set of common artists they have listened to:

$$J(u_1, u_2) = \frac{|Artists(u_1) \cap Artists(u_2)|}{|Artists(u_1) \cup Artists(u_2)|} \tag{11}$$

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

[1] The code is available at: https://github.com/pkouki/iui19.

$$Cos(u_1, u_2) = \frac{\sum\limits_{a \in \mathcal{A}} ListenCount(u_1, a) ListenCount(u_2, a)}{\sqrt{\sum\limits_{a \in \mathcal{A}} ListenCount(u_1, a)^2} \sqrt{\sum\limits_{a \in \mathcal{A}} ListenCount(u_2, a)^2}} \tag{12}$$

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 [31], 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 4 captures the intuition of item-based collaborative filtering methods, specifically that similar items will be listened by the same user. 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 11). Rule 5 is similar with the difference that we use last.fm's artist similarity, which is a proprietary score using CF and tag information. Rule 6 is inspired by the content-based collaborative filtering approaches and 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 11). Rule 7 is a simpler version of Rule 6 and captures the intuition that a user will likely listen to two artists sharing the same tag.

Rule 8 comes from the social-based filtering approaches and captures the intuition that friends listen to the same artists. Rule 9 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 10.

Our framework is flexible and can incorporate many other sources of available information by adding additional first-order rules. For instance, we can leverage demographic information by computing similarity neighborhood relationships in demographic feature space and employing the rules:

$\textsc{Country}(u_1, c) \wedge \textsc{Country}(u_2, c) \Rightarrow \textsc{SimUsers}_{Demo}(u_1, u_2)$

$\textsc{SimUsers}_{Demo}(u_1, u_2) \wedge \textsc{Listens}(u_1, a) \Rightarrow \textsc{Listens}(u_2, a)$.

In the beginning of this subsection we noted that the rules of the hybrid system are motivated by recommendation algorithms already in the literature, and provided a brief guide to how these rules map to existing approaches. However, by harnessing the power of a probabilistic programming language, our system is able to go beyond existing approaches that use a single algorithm and allow sophisticated knowledge fusion across recommenders. As an example, consider a matrix-factorization approach to user-based collaborative filtering. In such an approach, user-item ratings are used to learn a low-dimensional user representation that captures user similarity, and unseen ratings are predicted by using matrix multiplication. In HyPER, Rule 3 captures a similar user-based similarity to predict the ratings of unseen items using implication rather than matrix multiplication. However, an important contrast is that HyPER integrates these predictions across algorithms in a collective manner. So, for example, if a user U's predicted rating of an artist A is strongly supported by the tag-based similarity in Rule 7, $\textsc{Listens}(U, A)$ will have a high value, and this also increases the value for similar users via Rule 3, so that user U' who is similar to U will also have a higher value for $\textsc{Listens}(U', A)$. As a result, strong signals are propagated across recommendation algorithms, and probabilistic inference induces algorithms to agree on recommendations.

As discussed, the proposed hybrid music recommender system uses a basic set of the rules proposed in HyPER and extends this model by adding several rules to leverage additional information available in the music dataset (such as tags). The rules are defined by domain experts in the field. For example, for the music recommendation scenario, domain experts are aware of the fact that collaborative filtering, social filtering, and content-based filtering are very important signals for the recommendation engine, so they incorporate these signals as rules into the model. In this work, we rely solely on curated rules used in recommendation or defined by domain experts. In general, rules are very intuitive and of general purpose. This is the reason that the PSL framework has been applied in so many domains ranging from identifying disagreements to debate forums [39] to predicting student performance from online student behavior [45]. One interesting extension, beyond the scope of this work, would be to incorporate structure learning, i.e., the process of automatically learning the rules given a specific dataset. Several recent works describe structure learning for PSL models [14, 53], however generating appropriate explanations for novel, algorithmically-generated rules is beyond the scope of this work.

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 $\textsc{Listens}$ predicate. In other words, we find the most probable assignment to the unobserved variables ($\textsc{Listens}$) by performing joint inference over interdependent variables. After the inference completes for a user u, we select the $\textsc{Listen}(u, a)$ that scored in the top k positions. For each of the top k $\textsc{Listens}(u, a)$, 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 3.
- **Item-based CF** and **item-based last.fm**, with explanations similar to Table 1 (II-a, II-b), using the groundings of Rules 4 and 5 respectively.
- **Content-based Jaccard** and **content-based tags**, with explanations similar to Table 1 (III-a, III-b) using the groundings of Rules 6 and 7 respectively.
- **Social-based**, with explanations similar to Table 1 (IV) using the groundings of Rule 8.
- **Popularity-based**, with explanations similar to Table 1 (V) using the groundings of Rule 9.

As an example, assuming that for user Jen, the predicted value of the unobserved variable $\textsc{Listens}(Jen, U2)$ has the highest value among all other predicted values and, during inference, the following ground rules associated with $\textsc{Listens}(Jen, U2)$ were generated:

$$\begin{aligned}
\textsc{SimUsers}_{CF}(Jen, Aren) \wedge \textsc{Listens}(Aren, U2) &\Rightarrow \textsc{Listens}(Jen, U2)\\
\textsc{SimArtists}_{CF}(U2, ACDC) \wedge \textsc{Listens}(Jen, ACDC) &\Rightarrow \textsc{Listens}(Jen, U2)\\
\textsc{SimArtists}_{last.fm}(U2, Coldplay) \wedge \textsc{Listens}(Jen, Coldplay) &\Rightarrow \textsc{Listens}(Jen, U2)\\
\textsc{SimArtists}_{content}(U2, Beatles) \wedge \textsc{Listens}(Jen, Beatles) &\Rightarrow \textsc{Listens}(Jen, U2)\\
\textsc{HasTag}(U2, Rock) \wedge \textsc{HasTag}(Slayer, Rock) \wedge \textsc{Listens}(Jen, Slayer) &\Rightarrow \textsc{Listens}(Jen, U2)\\
\textsc{SimFriends}(Jen, Cindy) \wedge \textsc{Listens}(Cindy, U2) &\Rightarrow \textsc{Listens}(Jen, U2)\\
\textsc{PopularArtist}(U2) &\Rightarrow \textsc{Listens}(Jen, U2)
\end{aligned}$$

Note that in this example, many different rules, and correspondingly, recommendation styles, contribute to a single recommendation (Jen will listen to U2). The core research problem we discuss is determining how to explain this recommendation to the user Jen in the most effective manner. This requires choosing the appropriate explanation styles and presenting them in an informative manner.

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 evaluate our system on music recommendations for the last.fm website. We choose this platform because: i) it provides an API[2] offering convenient access to music data and ii) it contains a wide range of information that can be exploited by our hybrid model: user-artist interactions, user friendships, content information for artists (i.e., tags), and popular artists in the database. Last.fm builds a detailed profile of the users using "audioscrobbler", a system that records the tracks to which a user listens. Last.fm exposes two main API types, *Users* and *Artists*. The $User$ API provides access to user demographic information (e.g., country of origin), the user's top artists by listening frequency, and the user's friends. The $User$ API provides access to the user's top artists by listening frequency and the user's friends. Some of the methods are: $User.getInfo$ which returns general information about the user such as country of origin, $User.getTopArtists$ which returns the top artists in this user's profile, and $User.getFriends$ which returns the friends of this user. Similarly, the $Artist$ API provides the method $Artist.getSimilar$ that returns the most similar artists to an artist based on both collaborative information and tag information and the method $Arist.getTopTags$ that returns the top tags that this artist has been assigned to. Last.fm also offers general top-chart methods: the method $Chart.getTopKArtists$ returns the K artists with the highest number of listeners and playcounts in the database, while the method $Chart.getTopKTags$ returns the K tags that have appeared the highest number of times in the database. To integrate with last.fm's API we build a crawler using pylast[3] that allows us to collect information for each user in our study in real time.

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? Hybrid recommender systems (such as HyPER) that use a variety of information sources and signals to make predictions, can generate 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 [44] 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.

[2] https://www.last.fm/api
[3] https://github.com/pylast/pylast

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 [25] 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.

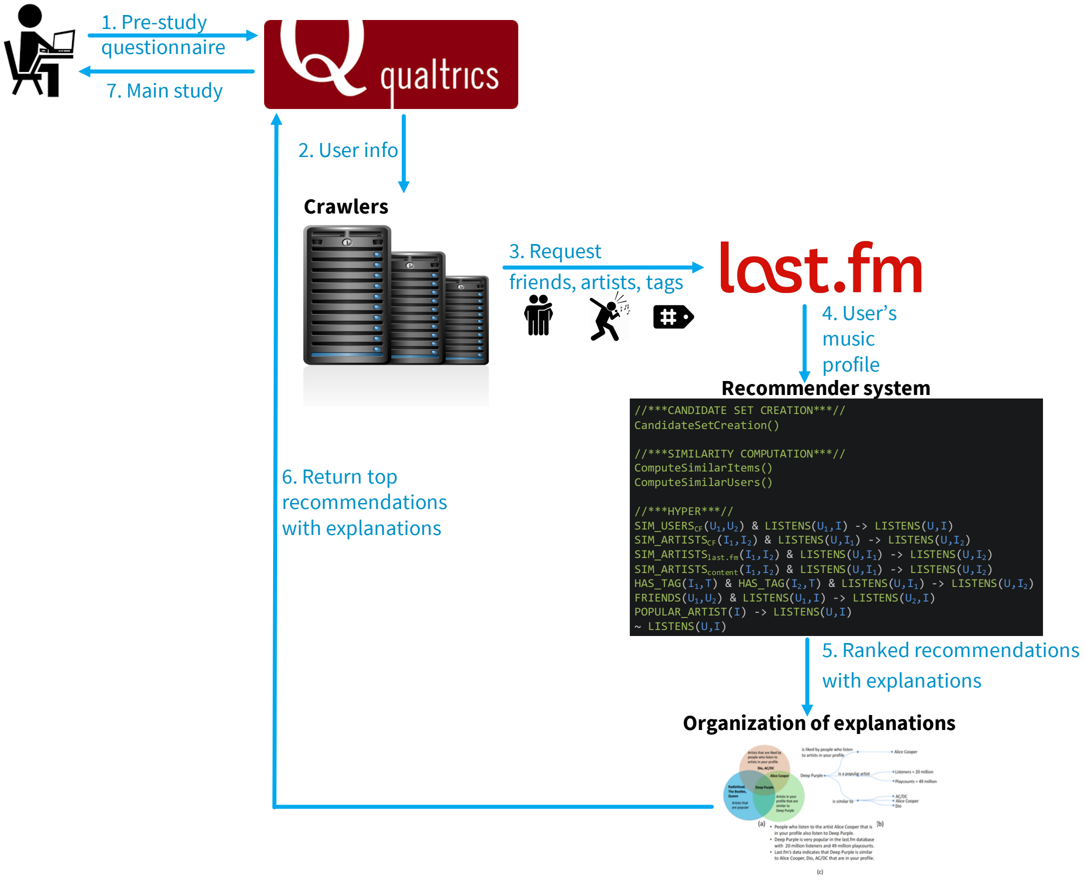
6 USER-STUDY DESIGN

We used the Amazon Mechanical Turk (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 provide their last.fm user name, and complete a pre-study questionnaire (this is the same for all participants). While the participant completed the questionnaire, we crawled the 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 asked a set of questions by following a methodology similar to Knijnenburg et al. [22]. Note that the recommended artists and the actual explanations were personalized to each participant, while the questions for evaluating recommendations and explanations were the same for all participants. We informed participants that part of the study will recommend artists. We acknowledged the fact that some recommendations might be unknown to the users, so we suggested that for these cases the users could click on the provided link (that accompanied each recommendation) that led to the last.fm artist profile to look into more information about the artist (e.g., description) or even go ahead and listen to some songs. Finally, we asked the participant for comments in free text. Next, we explain in detail the structure of the study for one participant which was divided into two basic phases. In Figure 1 we show graphically the steps performed during the user study.

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

In this first phase of the study, we informed the participants that, in order to participate, they needed to have a last.fm account with at least ten artists in their 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 a participant provided the 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 [38], visualization familiarity, and music familiarity. We additionally asked questions related to the five basic dimensions of personality, called the Big Five traits [44]. We adopted the abbreviated questionnaire by Gosling et al. [17] which is both brief and


1. Pre-study questionnaire
7. Main study
qualtrics
2. User info
Crawlers
3. Request friends, artists, tags
last.fm
4. User's music profile
Recommender system
6. Return top recommendations with explanations
5. Ranked recommendations with explanations
Organization of explanations
(a)
(b)
(c)

```
//***CANDIDATE SET CREATION***//
CandidateSetCreation()

//***SIMILARITY COMPUTATION***//
ComputeSimilarItems()
ComputeSimilarUsers()

//***HYPER***//
SIM_USERSCF(U1,U2) & LISTENS(U1,I) -> LISTENS(U,I)
SIM_ARTISTSCF(I1,I2) & LISTENS(U,I1) -> LISTENS(U,I2)
SIM_ARTISTSlast.fm(I1,I2) & LISTENS(U,I1) -> LISTENS(U,I2)
SIM_ARTISTScontent(I1,I2) & LISTENS(U,I1) -> LISTENS(U,I2)
HAS_TAG(I1,T) & HAS_TAG(I2,T) & LISTENS(U,I1) -> LISTENS(U,I2)
FRIENDS(U1,U2) & LISTENS(U1,I) -> LISTENS(U2,I)
POPULAR_ARTIST(I) -> LISTENS(U,I)
~ LISTENS(U,I)
```

Fig. 1. Illustration of the steps that take place during the study.

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 performed 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 Listens(u, a). The HyPER model makes predictions for **all** the unobserved 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 needed to 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) for each of the artists in the profile of the participant, we included the 20 most similar artists (based on the last.fm

Ease-of-Satisfaction (α = 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 (α = 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
Music Familiarity		
Compared to my peers I am an expert in music.	0.55	
I am a music lover.	0.78	
Compared to my peers I listen to music a lot.	0.52	
I closely follow artists/bands that I like.	0.39	
I am not into music.	0.26	
Personality - I see myself as...	Trait	
Extroverted, enthusiastic. Reserved, quiet.	Extroversion	
Dependable, self-disciplined. Disorganized, careless.	Dependability	
Open to new experiences, complex. Conventional, uncreative.	Openness	
Calm, emotionally stable. Anxious, easily upset.	Neuroticism	
Sympathetic, warm. 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.

similarity), (ii) for each of the friends in the profile of the participant we included the 20 top artists in this friend's profile and for each of these artists we got the top 20 most similar artists (again based on the last.fm similarity), and (iii) the top 1,000 artists in the last.fm database. For example, assume that a participant had 10 artists in his/her profile and 15 friends where each friend had 10 artists in her/his profile. When applying the above process, the candidate set contained 10*20=200 artists from step (i), 15 friends * 10 artists + 15 friends * 10 artists/friend *20 similar artists/ friend's artist = 3,150 artists from step (ii) and 1,000 artists from step (iii) which resulted in at most 4,350 artists in total (in cases with overlapping artists this number was smaller)

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 3), (ii) similarities between the participant and her friends (used by Rule 8), and (iii) content and CF similarities for the artists in the participant's candidate set (used by Rules 4 and 6). 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 [31] 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 would listen to artist a ($a \in \mathcal{A}$) (i.e., predict the values of the unobserved predicates $\textsc{Listens}(u, a)$). At the end of inference we picked the predictions for the predicates $\textsc{Listens}(u, a)$ that scored the highest.

Organize the explanations: To organize the explanations, we grouped multiple explanations of the same style together [25]. For example, if there were three groundings of the Rule 3 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 picked the groundings that involved 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 asked questions about the accuracy and novelty of the recommendations and questions related to the explanations provided with a focus on the persuasiveness aspect.

In particular, we showed the first best recommendation to address question 1, second best to address question 2, and the third best to address question 3 without introducing randomization. This was a design choice that was decided after carefully considering our goals in this work. The most important reason that led us to this decision is that introducing randomization would make the confounding effect harder to measure. The quality of the first recommendation might be excellent while the quality of the third recommendation might not be so good. For the case that the quality of the recommendation affects the explanation styles, the ranking of the styles, or the visuals, it would not be possible to compare answers given by different users when the quality of the recommendation varied, because the answers might have also varied when varying the quality. 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. An important note here is that the recommendation that was shown to the user was generated using the hybrid model described in Section 3.2. This indicates that multiple signals were used to come up with this recommendation, and as a result, multiple explanations were produced (as explained in Section 3.3). However, since in this part the goal is to study user preferences for different explanation styles, we did not show all explanations at once. Instead, we showed the exact same recommendation with a single explanation style generated during this prediction. We kept doing this for all other explanation styles (the maximum number of explanation styles generated was seven). With this task we tested the following hypotheses:

- H_1: Explanation style significantly correlates with perceived persuasiveness.
- H_2: Persuasiveness of explanation style is dependent on different personality types. This is a meta-hypothesis that consists of 70 different hypotheses. This is because we had responses from ten different personality questions and our framework supported seven different explanation styles. As a result, we had to test for an effect of each of the ten personality types on the seven different explanation styles.

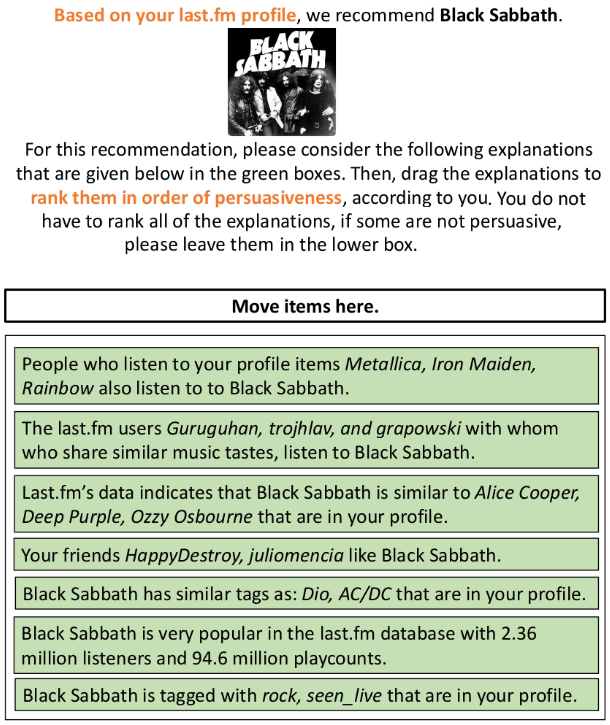
Task for research question 2: We showed the participant an artist profile for the second highest ranked artist as predicted by the HyPER model. We did not provide an explanation and asked the same questions related to perceived accuracy and novelty. Next, we showed the participant all the explanation 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 found uninteresting. Again, we randomized the initial order of the styles. Figure 2 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: Preferred number of explanations depends on different personality types. Similarly to H_2 this is a meta hypothesis that consists of 10 different hypotheses since we had to test for an effect of each of the ten personality types on the preferred number of explanations.

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 [25], which showed that Venn diagrams significantly outperform more complex visualizations such as concentric circles and 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 3 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 showed the different formats. With this task we tested the following hypothesis:

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

In this online study, we also asked the users to evaluate recommendations coming from a random recommender. In more detail, we recommended three random artists to the participant, i.e., we selected three artists from the candidate set $\mathcal{A}$ at random. For each artist, we showed their official picture as well as the link to the artist's last.fm account, but did not provide an explanation for the recommendation. Just as in the case of the recommendations coming from



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. 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 *Guruguhan, trojhlav, and grapowski* 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 *HappyDestroy, juliomencia* 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.

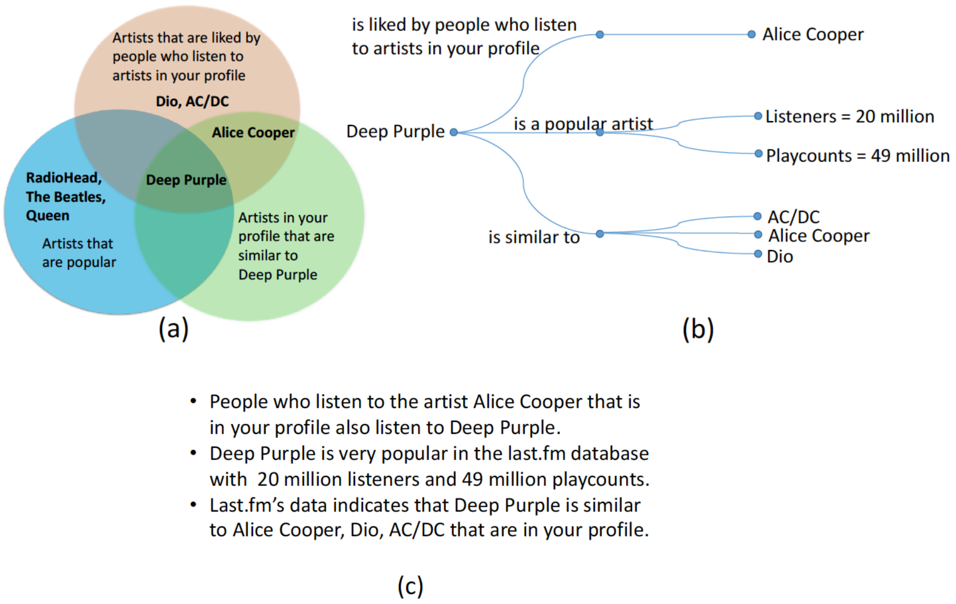
Fig. 2. Example of the ranking question (task for research question 2 of the study) for the recommended artist "Black Sabbath".

HyPER, for each of the randomly recommended artists, we asked the participants to rate the accuracy and novelty of the recommendation using questions in Table 3 (under "Perceived Accuracy" and "Perceived Novelty"). We compared these responses to those for HyPER later in the experimental results. Comparing to a random recommender system helped us: (1) do a sanity check to show that the recommendations provided by our framework were far better than a very weak recommender system and (2) do a sanity check for users' perspective. Recent literature [38] stated that it is effective to take a random model as a perception check of the users participating in the study. Obviously, the comparison with a random recommender system makes no claim about the overall accuracy of HyPER, since this is not within the scope of this work. Our aim in this paper is to measure personalized explanations, not the effectiveness of our recommendation algorithm. This has been studied in our previous paper [23] which extensively compared the performance of the recommendations provided by HyPER with other state-of-the-art systems.

In the middle of the study, we also asked the satisficing question: *"Please answer "somewhat not persuasive" to this question"*. This question, which was 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 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


Artists that are liked by people who listen to artists in your profile
Dio, AC/DC
Alice Cooper
RadioHead, The Beatles, Queen
Deep Purple
Artists that are popular
Artists in your profile that are similar to Deep Purple
(a)
is liked by people who listen to artists in your profile
Alice Cooper
Deep Purple
is a popular artist
Listeners = 20 million
Playcounts = 49 million
is similar to
AC/DC
Alice Cooper
Dio
(b)
(c)

- People who listen to the artist Alice Cooper that is in your profile also listen to Deep Purple.
- Deep Purple is very popular in the last.fm database with 20 million listeners and 49 million playcounts.
- Last.fm's data indicates that Deep Purple is similar to Alice Cooper, Dio, AC/DC that are in your profile.

Fig. 3. Example of the different explanation formats for the same recommended artist "Deep Purple" (task for the research question 3 of the study). (a) Venn diagrams, (b) static cluster dendrograms, (c) textual. We also showed interactive cluster dendrograms which were the same as static (b) with the difference that the participant could interact with the blue bullets (open or close them).

Perceived Accuracy (α = 0.96)	R^2	*Est.*
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 (α = 0.94)	R^2	*Est.*
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) (α = 0.93)	R^2	*Est.*
(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.

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 [20]. 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. Having to filter out only a small percentage of the Turkers had to do with the fact that we took several steps to safeguard against unreliable Turkers before startung the study. In order for someone to participate in our study, there was a prerequisite that she had to provide her last.fm account with at least 10 artists in her profile. Multiple Turkers attempted to participate in our study by providing a last.fm username that did not exist; we excluded these users. In other cases, the Turkers did create a new last.fm account but then they did not add any artists to their profile. Such cases were also excluded. In general, our system was able to detect all cases where either the username was invalid or the cases where there were less than 10 artists in the last.fm profile of the Turker. As a result, these Turkers were not allowed to continue with the study. Only when the prerequisites were met was the Turker able to fill in the main questionnaire. We believe that these initial steps were very important and filtered out many of the Turkers that were most inclined to give unreliable answers. In addition, we required that Turkers had at least 50 approved hits in their account before participating; this was another quality filter which indicated that these participants had successfully completed a sufficient number of other studies. We also restricted the location of users to US-only, since last.fm is not popular in all countries. And finally, we asked the users to participate to the study only if they were native English speakers.

Next, we computed the mean average time that it took for the participants to finish the study. This was 53 minutes with a standard deviation of 29 minutes. The median was 46 minutes. The highest value was 118 minutes and the lowest 10 minutes. This duration includes (a) the time that it took for the crawler to get the user's data (which overlapped with the time the user filled in the pre-study questionnaire) and (b) the time needed by the user to fill in the main study questionnaire. The time the crawler needed to get the user's data was highly dependent on the richness of the information of that particular user and her friends. For example, for a participant that had only 10 artists in the profile, the duration of crawling data was significantly lower compared to the duration that it took to crawl the data of a user with 20 artists and 20 friends where each friend had also a fair number of artists in her profile. As a result, out of the 53 minutes on average that it took for a participant to complete the study, a sizable fraction of the user time was spent waiting for the crawler. At the beginning of the study, we notified the users that part of the study would require waiting for recommendations to be generated. Also, we believe that due to this waiting time, some of the participants left the study in order to do something else, and then returned to fill in the main questionnaire and this may justify the relatively high duration of the study. This also justifies the incentive of $3.

Additionally, we computed statistics about artists and friends in the participants' profiles. First, we found that the average number of artists per user profile was 13.9 with standard deviation 5.2. Since a prerequisite of the study was the participant to have at least 10 friends in her profile, the minimum number of artists per profile was 10. The maximum number of artists we crawled was set to 20. Regarding social connections, many users participated in the study without having any friends. We computed the average number of friends and this was 7.4 with standard deviation 12. 59% of the participants (i.e., 117 in number) did not have any friends in their profile. When computing the same statistics for participants that had at least one friend, we found that the average number of friends was 18 with a standard deviation of 12.7. The largest number of friends was 20 which was due to the upper limit we set for the crawler. One important note here is that if a user had no friends in her profile then rule 8 (reported in Section 3.2) about social connections was not instantiated and as a result, these users were not provided with any social explanation.

As part of a post analysis, we divided the participants based on two criteria: (1) the familiarity they had with last.fm before doing the study and (2) reported music familiarity. In particular, for (1) we used the date that the participants created a last.fm account: 89 (45%) of the participants were considered "experienced" users who already had a last.fm account before starting the study, where 107 (55%) participants were "novices" who created a new last.fm account to participate to the study (we were not able to retrieve the date for two participants of the study). In each part of the study, we first report results for all the participants and then we report the same results separately for experienced and novice last.fm users.

Next, for (2), our initial goal was to use the answers to the five questions related to music familiarity and construct a single factor by performing confirmatory factor analysis (CFA). The reason that we asked five different questions related to music familiarity is that different users may interpret the same questions differently. CFA often reduces measurement error for questions posed to participants. Unfortunately, we found that we were not able to get a consistent fit between the five question items that were tested. R^2 measurements while trying to model all five questions in a single factor are presented in Table 2. Typically given these results, we would remove questions 4 and 5 that reported the worse values in terms of R^2 and re-test R^2 values, however we found that R^2 for question 1 was 0.44, R^2 for question 2 was 0.96, and R^2 for question 3 was 0.48. Essentially, this indicates that question 1 and question 3 were perceived somewhat differently by participants, but nearly all participants had answers that showed consistency either between question 1 and question 2 or question 2 and question 3. While keeping all three items would be an acceptable choice, we sacrificed some signal cleanliness for some internal reliability. Thus we dropped question 1 and question 3 and used question 2 as a standalone question item (*"I am a music lover"*). This means that, while we may have had to deal with some noise between music familiarity and other variables used in the study, at least the representation of music familiarity itself was consistent. As a result, we split the participants into two groups based on the answer they gave to the question: *"I am a music lover"*. Participants that replied *"agree"* and *"totally agree"* were considered as music experts, while participants that responded in the range of *"totally disagree"* to *"somewhat agree"* were considered as non-music experts. After the split, the group of music experts consisted of 97 (49%) participants while the groups of non-music experts consisted of 101 (51%) participants. In each part of the analysis below we study whether there is any statistically significant difference between the two groups.

	HyPER		**Random**	
Participants	**Accuracy (SD)**	**Novelty (SD)**	**Accuracy (SD)**	**Novelty (SD)**
All	5.64 (1.10)	2.10 (1.51)	3.53 (1.20)	5.52 (1.36)
Experienced	5.33 (1.14)	2.39 (1.67)	3.64 (1.20)	5.31 (1.43)
Novice	5.90 (1.02)	1.81 (1.23)	3.47 (1.17)	5.66 (1.26)

Table 4. Perceived accuracy and novelty for HyPER and random recommendations for (a) all the participants, (b) experienced last.fm users, and (c) novice last.fm users. Numbers in parenthesis indicate standard deviations.

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 [41]) 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 [9]. Each personality trait is determined by using an abbreviated measure adapted from [17]. To improve modeling of personality traits (which were not factored), we loaded a different latent variable on each response with a fixed value (1). Then, we freed the variance of each response and fixed 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 the subjective accuracy. We present all the results of means along with standard deviation values in Table 4. 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"*. Additionally, we computed how many of the artists recommended to the users belonged to the 1,000 most popular artists in last.fm. Out of the 198*3=594 artists which HyPER recommended to the participants, 47.8% (i.e., 284) belonged to the top 1,000 artists. As a result, more than half of the recommended artists were not popular artists. We underline that our method produces personalized recommendations without a focus on serendipity or novelty. As a result, if a user's data predict that she is most likely to listen to a popular artist then our recommender system will recommend this artist in the first place.

We followed the same process to evaluate the random recommendations and compare them with the hybrid recommendations. For each of the three random recommendations, we averaged the subjective accuracy. The random recommendations resulted in a mean accuracy of 3.53 out of 7 and mean novelty 5.5 out of 7. Thus, the improvement we get with HyPER in terms of perceived accuracy is 59.8%. At the same time, the improved accuracy provided by the hybrid recommendation was accompanied by a drop in novelty in comparison with random ($p < 0.001$, $S = 0.15$). As a result, the comparison of the hybrid recommendations with the random show what was expected: the recommendations coming from the HyPER framework are 59.8% more accurate than random recommendations, but this comes at a cost of less novel recommendations.

As discussed in Section 6.2 we used the first to third best recommendation to answer research questions from one to three respectively. Post-hoc, we studied how the perceived accuracy of the three provided recommendations varied for each user. Our goal was to check if the quality of the recommendations degraded quickly which might had an effect on the preference on different explanation styles, number of explanations, or explanation format. We computed that by looking into the answers to the accuracy question that reported the best R^2 value ("I like the recommended artist" with R^2=0.93). We found out that 41% of the users gave the exact same rating to the first and third recommendation, 44% of the users gave the same rating to the first and second recommendation, and 43% of the users gave the same rating to the second and third recommendation. Similarly, for 27% of the users the rating between the first and the third recommendation differed by one point (out of seven), for 31% of the users the rating between the first and the second recommendation differed by one point, and for 27% of the users the rating between the second and third recommendation differed by one point. In summary, for 70% of the users the rating between the first and third recommendation stayed either the same or changed by one point only, for 76% of the users the rating between the first and second recommendation stayed either the same or changed by one point only, and for 70% of the users the rating between the second and third recommendation stayed either the same or changed by one point only. This result showed that, in general, the users find that the provided recommendations have more or less the same accuracy, i.e., the quality of the recommendations does not degrade quickly.

Next, we computed the mean perceived accuracy and novelty separately for experienced and novice last.fm users. We found that the mean accuracy is 5.33 (out of 7) for experienced users and 5.90 for novice users. The novelty is 2.39 (out of 7) for experienced users and 1.8 for novice users. To compare perceived accuracy between novice and experienced users, we performed an unpaired t-test that shows the difference between these two groups is considered to be statistically significant ($p = 0.0003$). Comparing novelty between these two groups of users, we also found a statistically significant difference ($p = 0.006$). As a result, we conclude that the novice last.fm users find the recommendations from HyPER more accurate but less novel compared to the experienced last.fm users. The experienced users reported 3.64 perceived accuracy for the random recommendations while the novice users 3.47. Novelty was computed as 5.31 for experienced users and 5.67 for novice users for the random recommendations. The differences between the two groups related to the random recommendations were not statistically significant. We performed the same analysis for music and non-music experts. We did not find any statistically significant difference between the two groups.

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 [19]. In our analysis below, we use the accuracy and novelty as controlling variables and study their effect on explanations.

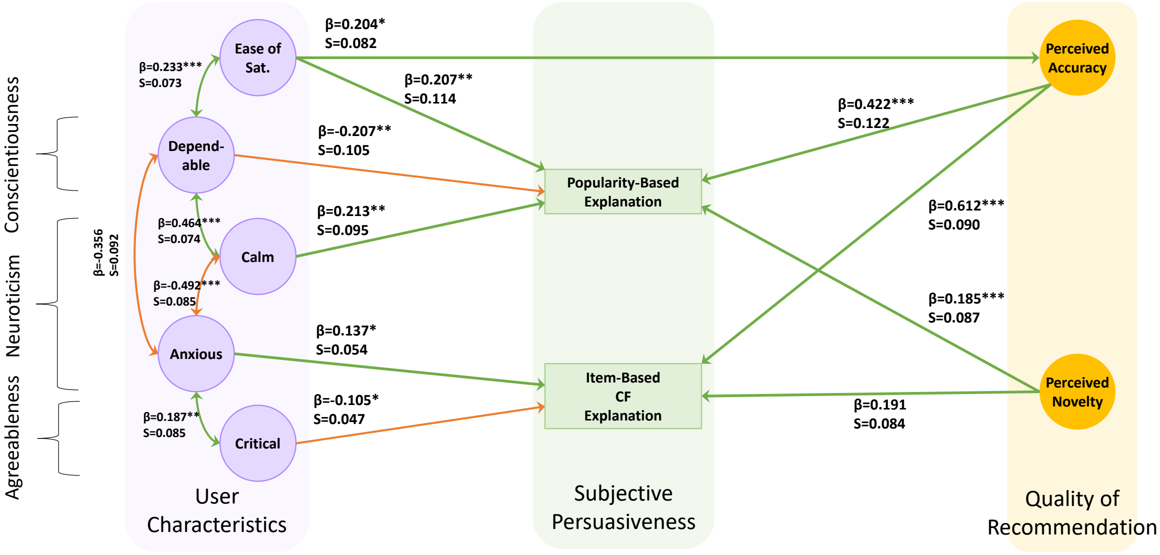


Fig. 4. An SEM explaining the role of personality in persuasiveness of explanation. Unidirectional arrows indicate regression, bidirectional arrows indicate co-variance; red arrows indicate a negative effect, green arrows indicate a positive effect; latent factors were scaled so β values indicate effect 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, $\chi^2(80) = 164.628$.

7.3 Preferences for explanation styles

In this section we answer the research question 1, i.e., *how does explanation persuasiveness vary with different explanation styles?*. To this end, we used the questions asked in Task 1 of the study to test for differences in persuasiveness when showing different explanation styles. Figure 5 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 [49] 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 ($\forall p < 0.001$). No significant improvement was found for item-based last.fm over item-based CF, or content-based Jaccard. According to additional analysis, most recommendations had almost all explanation styles. The total number of explanation styles that were shown to the participants can be found in the x axis of Figure 5. In particular, out of 198 recommendations the model showed: (a) 198 explanations of item-based CF last.fm style, (b) 198 explanations of content-based Jaccard style, (c) 198 explanations of content-based tags style, (d) 195 explanations for user-based style, (e) 195 explanations of item-based CF style, (f) 177 explanations of popularity-based style, and (e) 64 explanations of social style. The number of social explanations provided to the users was relatively small which is due to the fact that a large number of users participated in the study without having social connections in last.fm.

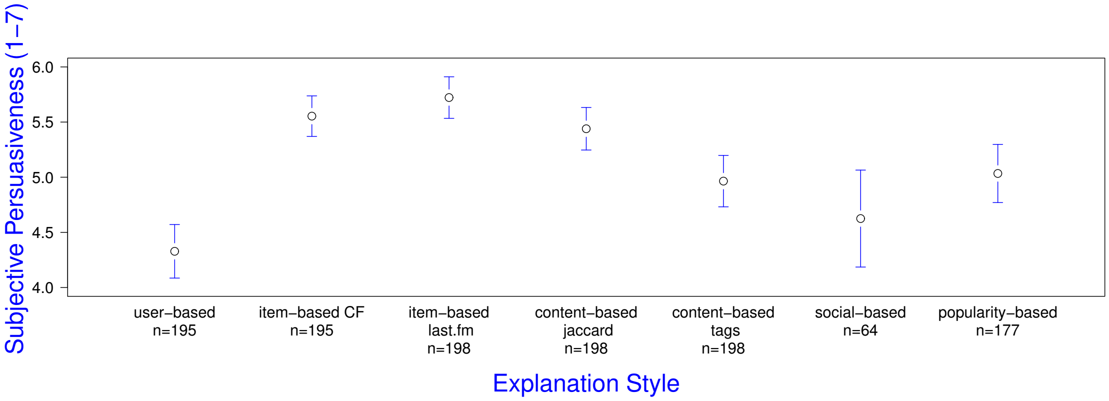


Fig. 5. All last.fm users: Mean subjective persuasiveness for each style of explanation, taken on a Likert Scale (1-7).

To test the significance of personality traits in the persuasiveness of an explanation, we conducted an exploratory structural equation modeling (SEM) [50] analysis. It is well known that people may change their ratings of items based on user experience or persuasive explanations [18], so we accounted for this effect by controlling for the accuracy/novelty of each recommendation and the participant's self-reported ease of satisfaction [38]. 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 [12].

Figure 4 shows the results from the exploratory analysis for all last.fm users. Of the ten personality traits, only four (dependable, calm, anxious, critical) were shown to have a significant bearing on persuasiveness of two explanation

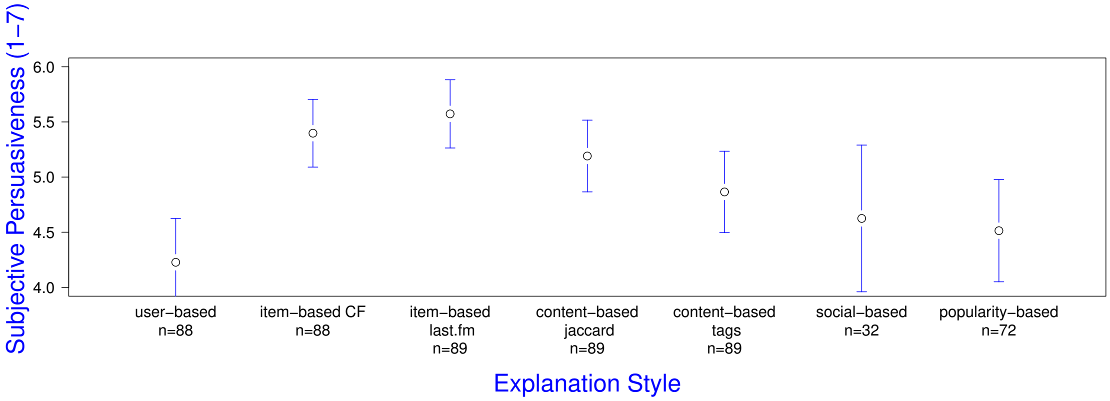


Fig. 6. Experienced last.fm users: Mean subjective persuasiveness for each style of explanation, taken on a Likert Scale (1-7).

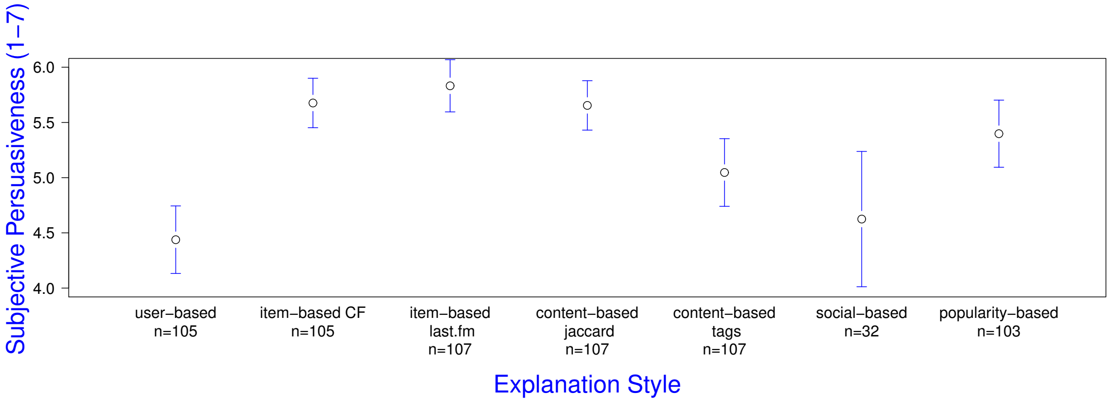


Fig. 7. Novice last.fm users: Mean subjective persuasiveness for each style of explanation, taken on a Likert Scale (1-7).

styles (popularity-based and item-based). Any variable with no significant effect was removed from the final model, and this is the reason that only two explanation styles and four personality types appear in this figure. 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, out of the 70 hypotheses that constructed the meta-hypothesis H_2, the data supports that we can accept only two of those: (1) persuasiveness of popularity-based explanations is dependent on the personality type *calm* and (2) persuasiveness of item-based CF explanations is dependent on the personality type *anxious*. We discuss these finding in detail in Section 8. Finally, the effect sizes of perceived accuracy appeared to be double that of perceived novelty and any personality-based effect.

Furthermore, we report the same graphs for experienced last.fm users in Figure 6 and novice last.fm users in Figure 7. To compute whether there are any statistically significant differences between the two subgroups regarding the preference to the explanation styles, we performed unpaired t-tests. Our analysis showed that novice last.fm users report statistically significantly higher preference for the content-based Jaccard and popularity-based styles compared to the experienced last.fm users. In more detail, the mean average for the experienced last.fm users for the content-based Jaccard explanation style was 5.19 (SD = 1.54) while for the novice users the mean was 5.65 (SD = 1.17) so the difference was considered to be statistically significant with $p = 0.0178$. Similarly, the mean average for the experienced last.fm users for the popularity-based explanation style was 4.51 (SD = 1.97) while for the novice users the mean was 5.40 (SD = 1.56) so the difference was considered to be very statistically significant with $p = 0.0011$. Again, we performed the same analysis for music and non-music experts but did not find any statistically significant difference between the two groups. We omit to show the values for each group since they are very close to each other.

7.4 Preferred number of explanation styles

In this section we answer the research question 2, i.e., *what is the ideal number of explanation styles?*. To this end, we analyzed the orderings given by the participants in the ranking questions (Task 2 of the study). First, we note that if the rankings were treated as ratings (1^{st} position = 7 points, 2^{nd} position = 6 points, etc.), each explanation style had the same relative score as shown in Figure 5 (this serves as a second level of validation for explanation preferences). Second, the mean number of explanation styles ranked was 2.61 (standard deviation (SD) = 2.69). However, we found that 39.4% of participants left explanation styles in the bottom box without ranking them.

Careful inspection of the data showed that most of the participants that did not rank any style was due to technical issues with the library used to provide the drag-and-drop functionality. These issues reflect our oversight to test the library in all browser and platform combinations during development, or require participants to have a minimal browser and OS version. Since these same participants rated recommendations prior to this phase of the experiment and rated visualizations after this phase of the experiment, we believe the interpretation that technical issues prevented them from responding is more likely than the interpretation that no explanations were meaningful. Additionally, we carefully inspected the comments we received from free text that confirm our claim. Here are some of the comments: *"The question: Please rate the following by moving the items - Did NOT work. I couldn't move, drag n drop any of the selections."*, *"There is a glitch in one section of the survey. I could not rank items since dragging and dropping the items in the top box did not work. Other than that, the survey was fine."*, *"It wouldn't let me drag and reorder the different descriptions."*, *"The move boxes section didn't work."*, *"It wouldn't let me drag and reorder the different descriptions."* This feature was not used in any other part of our study and thus there are no other issues that we are aware of (this also agrees with the feedback we got from the participants).

After removing these participants, the mean number of explanation styles was 4.32 (SD = 2.13). 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.

Additionally, we tested whether the preferred number of explanations depends on different personality types. As explained, 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 ($\beta = 0.212 **, S = 0.144$). As a result, we accept only two hypotheses out the meta-hypothesis H_4, i.e., (1) preferred number of explanations depends on the personality type *dependable* and (2) preferred number of explanations depends on the personality type *open*.

Finally, we computed the average number of explanations preferred for experienced and novice last.fm users. We found that the mean number of explanation styles was 2.8 (SD = 2.78) for experienced users and 2.5 (SD = 2.63) for new users. 40.7% of the experienced last.fm users did not rank any explanation style while 37.4% of the novice last.fm users did not rank any style. After removing those participants, the mean number of explanations was 4.73 (SD = 1.93) for experienced last.fm users and 4 (SD = 2.25) for novice last.fm users. None of those differences was considered to be statistically significant. Similarly, the differences between music experts and non-music experts were not considered statistically significant (we omit to report the actual values since they are very close to each other).

7.5 Textual vs. visual format

In this section, we answer the research question 3, i.e., *how does the explanation format affect user experience?*. To this end, 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 UXP questions reported in Table 3. Figure 8 plots the persuasiveness score which reported the best R^2 value (*"This explanation for the recommendation is convincing"*), for each visual/textual format. 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 ($\beta = 0.151*, S = 0.077$). Our analysis did not show any statistically significant difference between the static and interactive version of cluster dendrograms. Finally, we report the same graphs for experienced last.fm users in Figure 9 and novice last.fm users in Figure 10. Performing unpaired t-tests showed no statistically significant differences between the two subgroups. Again, we did not find any differences between music and non-music experts (we omit to show the graphs because they look almost identical).

7.6 User comments in free text

As discussed, at the end of the study we asked users to provide comments in free text. The comments we received from the users that are related to the different explanation styles are the following:

- *"I liked when the categories like pop, rock or alternative are included in the recommendation. That would help with the choice to click on one."* (i.e., in favor of the content-based tags explanation style)
- *"I don't think the high volume of listening is a good factor since many artists that I like are not high profile big bands, for instance, Iron Butterfly, and 10 years After. These groups are not as popular as Led Zeppelin, but fit my music tastes."* (i.e., against the popularity-based explanation)
- *"I might not be in the majority, but my tastes are very eclectic. As such, I feel that just because I listen to Band A and Band B doesn't mean others will like both. Since my tastes aren't well matched I'm less likely to be convinced when a platform says that "User A and myself both like Band A, and User A likes Band B, I'll like Band B.""* (i.e., against the user-based and social-based explanation)

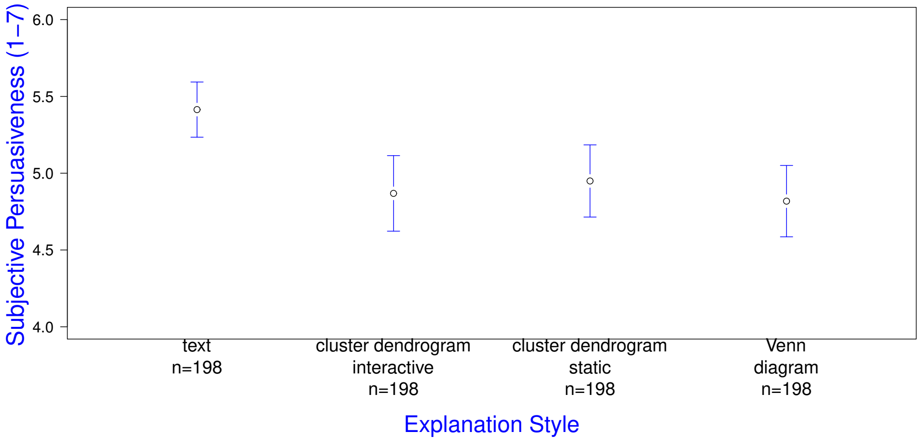


Fig. 8. All last.fm users: Mean subjective persuasiveness for each format of explanation, taken on a Likert Scale (1-7).

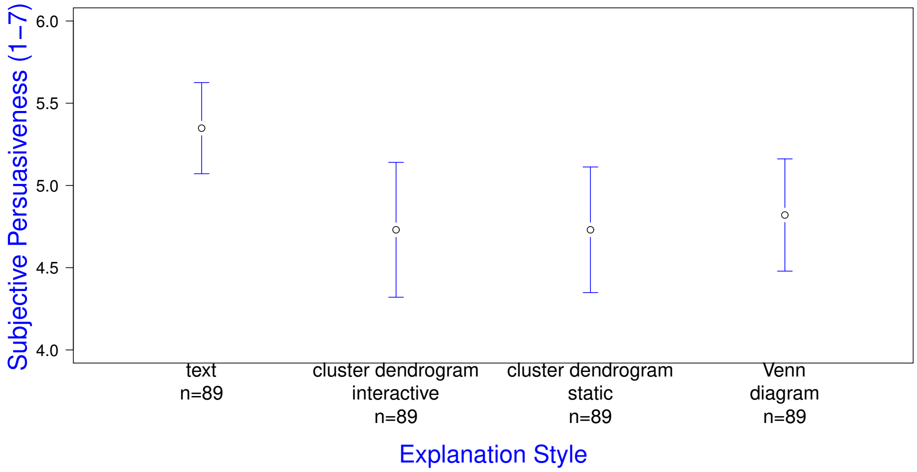


Fig. 9. Experienced last.fm users: Mean subjective persuasiveness for each format of explanation, taken on a Likert Scale (1-7).

- *"The tagged data doesn't really work for me. 'Seen live' tag does nothing for me."* (i.e., against the content-based tag explanation style)

We received two comments in free text from the participants that are related to the number and detail of explanations. The first comment agrees with the result of the statistical analysis, while the second one is from a user interested in receiving more detailed information. In more detail here are the comments:

- *"I think that the recommendations were on point. It is especially important that people aren't overwhelmed with information, and I think the simple graphs provide the best presentation to the end user."*

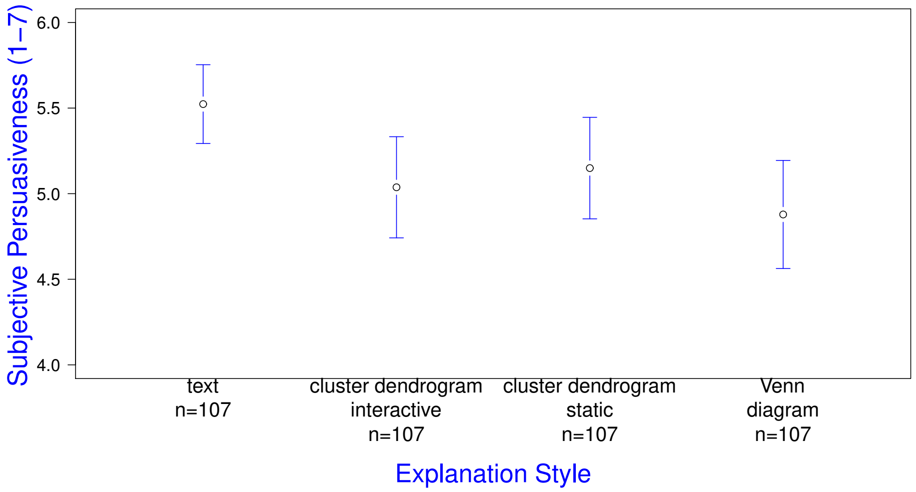


Fig. 10. Novice last.fm users: Mean subjective persuasiveness for each format of explanation, taken on a Likert Scale (1-7).

- *"It would be more persuasive if it used more comparisons, i.e this band is similar to these other 5 bands or people who liked this band also liked)"*

Users provided interesting comments related to the preferred format they would like explanations to be presented to them. As expected many times user preferences conflict, e.g., some users had strong preferences for Venn diagrams and other comments suggested that Venn diagrams should be avoided. Here we report the most insightful comments:

- *"The recommendations where it branches out to other info and facts was my favorite." (i.e., in favor of cluster dendrograms interactive)*
- *"Avoid venn diagrams.....they are never appealing." (i.e., against Venn diagrams)*
- *"I liked the Venn diagram best for this type of information! It was fairly unique." (i.e., in favor of Venn diagrams)*
- *"I think that the explanation tree is very unique and helpful." (i.e., in favor of cluster dendrograms)*
- *"I liked the trees presentation - it looked good and was simple and made me feel involved in the decision-making process." (i.e., cluster dendrograms, probably the interactive version)*
- *"I really like the diagrams. Especially the overlapping circles. The branching tree was good too." (i.e., in favor of the Veen diagrams and cluster dendrograms)*
- *"I feel that the diagrams were overall better options to persuade people and myself of the music recommendations."*
- *"I think the simple graphs provide the best presentation to the end user." (i.e., in favor of the graph-based formats)*
- *"I prefer text best rather than diagrams for recommendations." (i.e., in favor of textual explanations)*
- *"I feel that the diagrams were overall better options to persuade people and myself of the music recommendations." (i.e., in favor of graph-based explanations)*
- *"The last option of recommendation system (tree with dots)" (i.e., interactive cluster dendrograms) "was similar to the first," (i.e., cluster dendrograms static) "but I didn't like it as much because you had to click each dot to get the information, where*

it just presented to you quickly in the first option. The Venn diagram was pretty confusing. The text descriptions work, but don't carry the weight of having the graphical recommendations.' (i.e., in favor of the static cluster dendrograms)

Finally, we received interesting feedback about our study in free text from some of the users. Most of the feedback we received was encouraging for our study and sometimes encouraging for the last.fm music platform:

- *"Seems there is too much algorithms and data analysis designed in recommending music and artists to users. Music is inherently a subjective thing so I'm not sure I think this is the best method."*
- *"I think using a faster website than last.fm and making the survey font bigger will help a lot."*
- *"Fun and interesting survey, thank you for the opportunity."*
- *"It ran smoothly."*
- *"Good survey."*
- *"Interesting study."*
- *"Fun!"*
- *"Thank you for thanks for introducing me to an artist who I now like that I was unaware about.ucing me to this platform, I will get many enjoyable listening experiences with it and the best part is that I will be assisted in finding artists that I have forgotten about."*
- *"Very interesting."*
- *"Very interesting tool to predict music tastes. I can see where it would work very well. I can also see where it might not work so well."*
- *"Seems like a cool idea."*
- *"Nice survey."*
- *"I learned new things."*
- *"Interesting. Had fun doing this!"*
- *"Everything is good."*
- *"I actually have never heard of this music site and I enjoy it very much. I plan on using it in the future as well, thank you."*
- *"Fun survey thank you." "I feel that during the music recommendation process that I should have had an option to what some of the artist a test? because some of them the music i would never listen to.ize my music tastes myself."*
- *"I enjoyed the site. It is a new site for me to interact with in the future."*
- *"Great survey. I enjoyed it!"*
- *"Great, interesting study! Thank you"*
- *"Good and smart."*
- *"I thought it was interesting and insightful to learn about how the music is related to other music."*
- *"Cool study. i like the new ways to find new artists."*
- *"I don't understand how your website works and I feel old when I use it."*
- *"This was a very interesting approach. I would be interested in trying it out in the future."*
- *"I really like the way the suggestions are laid out in this survey."*
- *"Thank you! I enjoyed the study with no problems except that I had to submit my MTurk HIT early due to PC issues."*
- *"Good luck. Glad to be a part of last.fm now."*
- *"I thought the live recommendation system was pretty cool. I'd probably use a tool like that to find new music. I also didn't experience any technical issues."*
- *"Honestly, the interactive system is great! I'd love to see it in the wild."*

- *"Wow! This is very cool! I hope to see what comes out of this researchi think that it is a good platform to find music and people with the same taste in music you guys!"*
- *"I think that it is a good platform to find music and people with the same taste in music."*
- *"I enjoyed the artists presented to me and the ways of discovering new ones."*
- *"It was fun and engaging!"*
- *"I thought this was a cool study. The presentation was nice."*
- *"I enjoyed taking this survey, the questions were clear and I enjoyed some of the music."*
- *"This is a very good way of recommending new music/and or stuff I like. Finding new music is always a bit time consuming and this could save me a lot of time."*

8 DISCUSSION

In this work, we implemented a personalized hybrid recommender system that combines multiple sources of information to generate recommendations with a variety of explanation styles in 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. 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 [16, 18] has shown that users can evaluate content-based explanations with precision and that a content-based interface is preferable. The findings in this work reinforce the idea 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. Receptivity to popular items makes sense for these users, as they also scored highly in extroversion and sympathy. An additional explanation is that calm individuals do not try to find their own nuance and are happy to go with the flow. Moreover, neurotic individuals try to differentiate themselves from the others and tend to have a desire to cover their own needs in a way that is different from the opinion of the crowd. We recommend further research and user studies to explain the exact reasons of the above correlations. Participants that identified as dependable did not have any preference for the popularity-based explanation, potentially suggesting they are affected less by popular opinion. Likewise, neurotic participants (who were also likely to be introverted and reserved), showed a slight preference for item-based CF explanations, which surface patterns of particular tastes shared with others. Critical (disagreeable) users did not find the item-based CF explanation any better than average.

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 prefered to see no explanation with a recommendation, which we believe is an 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 and disorganization. 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 better. 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 were 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. [36] proposed Venn diagrams to implement interactive interfaces and showed that they are more engaging to the users when compared to non-controllable ranked lists. We draw a distinction between such interactive visualizations that support a broader capability for data exploration which have a rich history of development in the information visualization community, and our work that evaluates primarily non-interactive visualizations solely on their persuasiveness in a less task-focused environment. In another study [25], 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.

Experienced and novice last.fm users differ in three significant ways. As discussed, 45% of the users participated in the study already had a last.fm account, while 55% of the users created a last.fm account to participate to our study. A variety of unpaired t-tests and 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 ($B = 0.576 * **$) but lower perceived novelty, (2) preferred the content-based Jaccard explanations more ($B = 0.544*$), and (3) 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 had already explored the online space of music offerings to a greater extent and as a result they were harder to satisfy when compared to novice users. On the contrary, novice last.fm users that experience the last.fm music platform for the first time may be satisfied from its capabilities. A possible explanation for the fact that novice users reported lower novelty is the fact that those users likely had less complete profiles and listening histories and as a result the recommended artists were those the users were already familiar with but had not yet added to their profile.

The aforementioned findings may suggest that an explainable music recommender system should treat experienced users differently than novice users, or support a new breed of recommender systems that adaptively update suggestions as the user grows more familiar with a domain.

Music and non-music experts do not report different preferences. 49% of participants were considered as music experts while 51% of the participants considered as non-music experts. These two groups did not report any statistically significant difference in terms of perceived accuracy, novelty, and explanation preferences. This result suggests that an explainable music recommender system does not have to treat differently users based on their music expertise. Again, this result may seem to contradict with work on explainable music recommender systems [34, 28] , but like before, the results are not directly comparable. In detail, Oramas et al. [34] showed that effectiveness of content-based explanations depends on the the music education of the users, Millecamp et al. [28] showed that users with a higher musical sophistication tend to listen less songs when evaluating them, while we showed that music expertise is not a factor that affects the preference of users in different explanation styles.

9 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. Here is a list of the limitations of this work along with ideas about how these can be addressed in the future:

- Study the optimal number of explanations to display: In the music domain we concluded that people prefer to see three to four explanations. A basic assumption that we used to reach this conclusion is that all explanations provided were useful to the participant. However, there is also the case that some of the provided explanations were not useful to the participants and this is the reason that they did not include them in the ranking. In the future, we plan to further study if the ideal number of explanations to display is affected by the usefulness of the explanations provided. Additionally, we are interested in studying how the optimal number of explanations varies when varying the domain of the recommender system. For example, 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 to 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.
- Account for trust and privacy: Although our work provides a variety of explanation styles to the users to support the provided recommendations, we do not account for trust and privacy. Consider, for example, the social-based explanation of the form *"We recommend bar Crudo because your friend Cindy likes it"*. Although such an explanation seems benign it involves several potential issues regarding trust and privacy. First, in our example, we don't know how trustworthy *Cindy* is in order to use her specifically in explanations. Second, bar *Crudo* may be a bar attracting people of a particular sexual orientation and, in this case, we would be giving away *Cindy's* potentially sensitive information. In our future work we plan to identify how to protect the user from untrusted information and also how to protect users' privacy that may be compromised through explanations.
- Study whether explanations should participate in the process of ranking the recommendations: In the method we proposed here, the recommender system first ranks the recommendations based on evidence from the user's prior history and then generates the explanations. In our future work, we are interested in studying whether and under what circumstances it is beneficial that explanations participate in the process of ranking the recommendations. In particular, we plan to incorporate explanations and user preferences in the prediction process of the recommender system in order to improve the accuracy and transparency of recommendations. For example, consider a particular user that has expressed the preference on content-based explanations, but she does not appreciate explanations coming from friends. For such users, we can recommend by favoring items that involve many content-based explanations and, at the same time, refrain from favoring items recommended through social connections. The predicted recommendations will involve explanations that this user likes and omit explanations that she dislikes. The goal is to have a unified transparent recommender system that recommends not only based on user data but also on the quality, trust, and privacy of the explanations.
- Study in more depth the correlations between personality types and explanations: As part of this work, we studied the correlation between the personality types with the explanation styles and the number of preferred styles to show

(i.e., H_2 and H_4). To the best of our knowledge, this is the first work that studied these correlations. As discussed, to study the presence of any effects we used SEM where we performed a large number of regressions. Estimating sample sizes for SEM is a challenging problem. However, sample size estimation is usually performed before statistical testing in order to minimize cost and bias. For our study, we used 10 participants per question item. This is a commonly used rule of thumb. For SEM, we used the Lavaan package in R. This package does not assume that data is normal [4]. In addition all responses were based on a Likert scale from 1 to 7 which are assumed to be normally distributed. Our focus in this work was in (a) the overall model fit and (b) the statistical significance of any particular effect in the model, as well as the corresponding effect size. Note that the existence of (b) is not dependent on the goodness of (a). In other words, we can have real effects inside an otherwise poorly specified and ill-fitted model. While there is some room for improvement, the model reported in the paper achieved what is often considered as an acceptable model fit, and most of the interesting effects in the model that were discussed in the paper had p-values of <0.001. We consider the actual effects to have a somewhat more limited impact than overall model fit, since the model would need to be deployed in production in order to fully estimate its impact which is out of the scope of our work. We advise for more research in this direction that focuses on studying in more depth these correlations and explaining their existence or absence.

- Study the effect of different explanation styles beyond persuasiveness: explanations have many goals, varying from efficiency, to scrutability and trust [42]. Designing a user study to investigate all aspects of explanation is extraordinarily difficult, requires significant resources, and extremely careful experimental design. As a result, in this paper, we decided to focus only on one goal, i.e., persuasiveness, which aims at convincing users to try or buy [42]. The basic reason that led us to this decision is that because industrial applications benefit most when persuading a user to accept recommendations, so persuasion is an important practical concern. In the future we plan to explore the effect of different explanation styles to satisfaction, trust, and effectiveness.
- Study the effect of content-based explanations when using tags that have been generated by music experts: Content-based explanations in this study are generated using information about tags coming directly from last.fm. Manual inspection of the tags showed that last.fm contains a large number of tags, many of which do not make sense. To deal with the noisy tags, we used only the top 1,000 most popular tags in the last.fm platform and ignored others since they were probably not informative for the average user. In the list of the 1,000 tags we manually inspected all 1,000 tags and removed tags that contained offensive content (e.g., “nazi”, “silly”, “stupid”, “crap”). Content-based explanations using tags might be more successful in a different context where tags have been created by experts. At the same time, generating high quality tags requires music experts, is labor-intensive, and costly, and is out of the scope of this work. More research is needed to evaluate how tag-based explanations perform when varying the quality of the tags provided.
- Study the effect of social explanations in a network with richer social connections: In part of this study, more than half of the participants did not have any friend in their music profile, and as a result evaluating how persuasive social explanations are was hard. In the future, we plan to study the effect of social explanations in a network were social connections are not that sparse.
- Study the generalizability of our findings in different contexts and scenarios: In terms of this work, we performed one study conducted using the last.fm music platform with a fixed number and styles of explanations. The recommendations were produced using a specific hybrid recommender system [23] and the different visualization formats

[4] https://cran.r-project.org/web/packages/lavaan/lavaan.pdf

came from the results of our previous work [25]. We believe that there is a lot of exciting research in the area of hybrid explanations where many of the variables that were used as fixed in our study could be changed. For example, 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).

- Compare the performance of the HyPER framework in online and offline evaluation: As part of this study we compared the performance of the HyPER to a random recommender system that was obviously very weak. As discussed, the basic reason for doing this was to do a sanity check about recommendations produced by HyPER and about users' perspective. In our case, the results were as expected, i.e., users showed that they vastly preferred recommendations from a sophisticated recommender systems (HyPER) compared to the recommendations from a random system. However, if the opposite was true, i.e., users showed to prefer random recommendations or showed not to have any preference between the two recommender systems, then this would have been an indication that either (1) the recommendations produced by HyPER were poor or (2) the majority of participants did not pay attention to the study and as a result we could not have moved forward with data analysis. The comparison with a random recommender system makes no claim about the overall accuracy of HyPER, since this is not within the scope of this work, and was the focus of an earlier paper [23] that extensively compared the performance of the recommendations provided by HyPER with other state-of-the-art systems. However, an interesting future direction would be to evaluate the performance of the HyPER framework and state-of-the-art approaches in an online setting and compare those results with results from offline evaluation. This way we may be able to see whether results from objective metrics (such as perceived accuracy) given directly by users correlate with popular metrics used in offline evaluations (e.g., mean average precision). Possible candidates for comparison are systems that provide explainable recommendations in the music domain [29, 28, 33].
- Provide explanations in an interactive system that fosters exploration: The explanations provided in this work were in a static context (with the exception of interactive cluster dendrogram) where the user could not interact with the explainable recommender system and give feedback about her preferences. More than that, the current implementation did not allow for further exploration of the provided explanations by the interested user (again with the same exception). However, it might be the case that some users are interested into exploring explanations, so giving them the ability to explore all different kinds of explanations available might improve user experience. In the future, we plan to support an interactive recommender system that will use the observations from interactions with the user and adjust the explanations based on the user's preferences. In addition to that and inspired by IntersectionExplorer [10], we plan to implement a similar scalable visualization that allows users to explore recommendations as well as explanations.

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] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein. 2011. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends Machine Learning*, 3, 1, 1–122.

[9] D. Campbell and D. Fiske. 1959. Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, 56, 2.

[10] B. Cardoso, G. Sedraky, F. Gutierrez, D. Parra, P. Brusilovsky, and K. Verbert. 2019. Intersectionexplorer, a multi-perspective approach for exploring recommendations. *International Journal of Human-Computer Studies*, 121.

[11] 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).

[12] R. Cribbie. 2007. Multiplicity control in structural equation modeling. *Structural Equation Modeling*, 14, 1.

[13] Z. Dou, S. Hu, K. Chen, R. Song, and J. Wen. 2011. Multi-dimensional search result diversification. In *Proceedings of the 4th ACM International Conference on Web Search and Data Mining* (WSDM '11).

[14] V. Embar, D. Sridhar, G. Farnadi, and L. Getoor. 2018. Scalable structure learning for probabilistic soft logic. In *Workshop on Statistical Relational AI* ((StarAI 2018)).

[15] G. Friedrich and M. Zanker. 2017. A taxonomy for generating explanations in recommender systems. *AI Magazine*, 32, 3.

[16] 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.

[17] 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.

[18] J. Herlocker, J. Konstan, and J. Riedl. 2000. Explaining collaborative filtering recommendations. In *Proceedings of the Conference on Computer Supported Cooperative Work (CSCW '00).*

[19] J. Herlocker, J. Konstan, L. Terveen, and J. Riedl. 2004. Evaluating collaborative filtering recommender systems. *ACM Transactions on Information Systems.* (TOIS '04) 22, 1.

[20] 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).

[21] Y. Juan, Y. Zhuang, W. Chin, and C. Lin. 2016. Field-aware factorization machines for ctr prediction. In *Proceedings of the 10th ACM Conference on Recommender Systems* (RecSys '16).

[22] 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).

[23] 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).

[24] P. Kouki, J. Schaffer, J. Pujara, J. O'Donovan, and L. Getoor. 2019. Personalized explanations for hybrid recommender systems. In *Proceedings of the 24th International Conference on Intelligent User Interfaces* (IUI '19).

[25] 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).

[26] A. Lex, N. Gehlenborg, H. Strobelt, R. Vuillemot, and H. Pfister. 2014. Upset: visualization of intersecting sets. *IEEE Trans Vis Comput Graph*, 20, 12.

[27] Y. Lu, R. Dong, and B. Smyth. 2018. Coevolutionary recommendation model: mutual learning between ratings and reviews. In *Proceedings of the 2018 World Wide Web Conference* (WWW '18).

[28] M. Millecamp, N. N. Htun, C. Conati, and K. Verbert. 2019. To explain or not to explain: the effects of personal characteristics when explaining music recommendations. In *Proceedings of the 24th International Conference on Intelligent User Interfaces (IUI'19).*

[29] M. Millecamp, S. Naveed, K. Verbert, and J. Ziegler. 2019. To explain or not to explain: the effects of personal characteristics when explaining feature-based recommendations in different domains. In *Joint Workshop on Interfaces and Human Decision Making for Recommender Systems (IntRS'19).*

[30] K. Muhammad, A. Lawlor, and B. Smyth. 2016. On the use of opinionated explanations to rank and justify recommendations. In *FLAIRS '16*.

[31] X. Ning, C. Desrosiers, and G. Karypis. 2015. *A comprehensive survey of neighborhood based recommendation methods*. Recommender Systems Handbook, Second Edition, Springer US.

[32] 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.

[33] 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).

[34] 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.

[35] 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.

[36] 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).

[37] 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).*

[38] 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).*

[39] D. Sridhar, J. Foulds, M. Walker, B. Huang, and L. Getoor. 2015. Joint models of disagreement and stance in online debate. In *Proceedings of the 53th Annual Meeting of the Association for Computational Linguistics* (ACL'15).

[40] P. Symeonidis, A. Nanopoulos, and Y. Manolopoulos. 2009. Moviexplain: a recommender system with explanations. In *Proceedings of the 3rd ACM Conference on Recommender Systems* (RecSys '09).

[41] M. Tavakol and R. Dennick. 2011. Making sense of cronbach's alpha. *International Journal of Medical Education*, 2.

[42] N. Tintarev and J. Masthoff. 2015. *Designing and Evaluating Explanations for Recommender Systems*. Recommender Systems Handbook, Second Edition, Springer US.

[43] 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.

[44] M. Tkalcic and L. Chen. 2015. *Personality and Recommender Systems*. Recommender Systems Handbook, Second Edition, Springer US.

[45] S. Tomkins, A. Ramesh, and L. Getoor. 2016. Predicting post-test performance from online student behavior: a high school mooc case study. In *International Conference on Educational Data Mining* (EDM'16).

[46] C. Tsai and P. Brusilovsky. 2018. Beyond the ranked list: user-driven exploration and diversification of social recommendation. In *Proceedings of the 23rd International Conference on Intelligent User Interfaces* (IUI '18).

[47] C. Tsai and P. Brusilovsky. 2019. Explaining recommendations in an interactive hybrid social recommender. In *Proceedings of the 24th International Conference on Intelligent User Interfaces* (IUI '19).

[48] K. Tsukuda and M. Goto. 2019. Dualdiv: diversifying items and explanation styles in explainable hybrid recommendation. In *Proceedings of the 11th ACM Conference on Recommender Systems* (RecSys '19).

[49] J. Tukey. 1949. Comparing individual means in the analysis of variance. *Biometrics*.

[50] J. Ullman and P. Bentler. 2003. *Structural equation modeling*. Wiley Online Library.

[51] 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).

[52] 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).

[53] Y. Zhang and A. Ramesh. 2019. Learning interpretable relational structures of hinge-loss markov random fields. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, 6050–6056.