

The Effects of Controllability and Explainability in a Social Recommender System

Chun-Hua Tsai¹ · Peter Brusilovsky²

Received: 09 Sep 2019 / Accepted: DD Month YEAR

Abstract In recent years, researchers in the field of recommender systems have explored a range of advanced interfaces to improve user interactions with recommender systems. Some of the major research ideas explored in this new area include the explainability and controllability of recommendations. *Controllability* enables end users to participate in the recommendation process by providing various kinds of input. *Explainability* focuses on making the recommendation process and the reasons behind specific recommendation more clear to the users. While each of these approaches contributes to making traditional “black-box” recommendation more attractive and acceptable to end users, little is known about how these approaches work together. In this paper, we investigate the effects of adding user control and visual explanations in a specific context of an interactive hybrid social recommender system. We present *Relevance Tuner+*, a hybrid recommender system that allows the users to control the fusion of multiple recommender sources while also offering explanations of both the fusion process and each of the source recommendations. We also report the results of a controlled study (N=50) that explores the impact of controllability and explainability in this context.

Keywords User Interface · Hybrid Recommendation · Transparency · Controllability · Explainability

1 Introduction

The growth in artificial intelligence (AI) technology has advanced many human-facing applications. A *recommender system* is one of the best-known examples of applying the ideas of AI to a range of real-world applications. For instance, *social recommendations* (recommendations based on explicit or implicit social connections) have been widely adopted in many social media and e-commerce platforms.

E-mail: Chun-Hua Tsai
ctsai@psu.edu

¹ The Pennsylvania State University, State College, PA, USA

² University of Pittsburgh, Pittsburgh, PA, USA

Social recommendations attempt to filter out “irrelevant” information so users can reduce the efforts of decision making, such as purchasing an item online or following a new friend on social media. Social recommender systems typically use large-scale data from multiple data sources and process this data with complex intelligent inference methods; however, the “reasons” of offering specific recommendations usually stay in a *black box*, which frequently make the resulting recommendations “puzzling” to the users (Amershi et al., 2019). Herlocker et al. (2000) demonstrated that the user generally has little understanding about the mechanism behind these systems. In this situation, processing this output could produce user behavior that can be confusing, frustrating, or even dangerous in life-changing scenarios. Moreover, a lower-transparency system is known to negatively affect user satisfaction with recommender systems (Tintarev and Masthoff, 2015).

The problems of “black-box” recommender systems have been addressed in the new generation on *interactive recommender systems* (He et al., 2016). These systems attempted to make the recommendation process more transparent by visualizing some aspects of the process and offering the user some form of control over the process. Starting with the pioneering work on PeerChooser (O’Donovan et al., 2008), several attempts were made to produce more transparent and controllable recommender systems. In these systems, users were allowed to “influence” the presented recommendations by interacting with different types of visual interfaces. Several studies of interactive recommender systems demonstrated that users appreciate *controllability* in their interactions with the recommender systems (Knijnenburg et al., 2012; Harper et al., 2015; Parra and Brusilovsky, 2015; di Sciascio et al., 2018). It has also been shown that the visualization has helped users to understand how their actions can impact the system (Jannach et al., 2016), which contributes to the overall *inspectability* (Knijnenburg et al., 2012) and *transparency* of the recommendation process (Tintarev and Masthoff, 2015).

Another stream of research (Herlocker et al., 2000; Tintarev and Masthoff, 2015) argued that transparency and user satisfaction could be increased by providing proper *explanations* to help the user to understand the reasoning process of both the system and the generated recommendations. This approach aims to provide more details that may make the users better comprehend the reasons for receiving the recommendations. In many user-centered evaluations (Herlocker et al., 2000; Knijnenburg et al., 2012; Tintarev and Masthoff, 2012), the explanation positively contributes to the user experience; i.e., trust, understandability, and satisfaction (Gedikli et al., 2014; Tintarev and Masthoff, 2015). Self-explainable recommender systems have been proved to increase user perception of system transparency (Tintarev and Masthoff, 2012), overall levels of trust (Pu and Chen, 2007), and acceptance of the system suggestions (Knijnenburg et al., 2012).

Enhancing the explainability of recommender systems and AI-based systems in general has drawn more and more attention in the broader field of human-computer interaction (HCI) due to the newly initiated European Union’s General Data Protection Regulation (GDPR), which requires the owner of any data-driven application to maintain a “right to the explanation” of algorithmic decisions (Eiband et al., 2018). The ongoing advance in AI techniques generates tremendous opportunities as well as challenges for the user interface and interaction design, which have been extensively discussed and debated in the HCI field (Amershi et al., 2019). For instance, how can we design an effective explanation interface? How can the

explanations affect the recommender system controllability? How does the user adopt these explanations in their own decision-making and information seeking processes?

While controllability and explainability demonstrated their value in making traditional “black-box” recommendation more transparent, attractive, and acceptable to the end users, little is known about how these approaches work together. This paper presents our exploration of both controllability and explainability in the context of a hybrid social recommender system.

Hybrid recommender systems attempt to improve the quality of recommendations by engaging several recommendation sources or approaches (Burke, 2007). While hybrid recommender systems are known for their performance, their sophisticated computational processes might be even more confusing to end users than single-source approaches. Even the simplest *parallel hybridization* approach, which fuses multiple recommendation sources with different weights, might appear opaque to end users (Burke, 2007). Usually, the optimal fusion weights are trained or learned using ground truth data to optimize for the best overall performance. The result might be optimal in general, but neither truly useful in a specific context nor easy to comprehend.

In many real-world scenarios, a hybrid social recommender system can be applied in response to different user needs (Tsai and Brusilovsky, 2018) and even by different stakeholders (Abdollahpouri et al., 2019). For example, in an event-based social recommender (Brusilovsky et al., 2016), the event attendees may see recommendations of people to meet for a range of reasons; for example, a conference attendee may wish to re-connect with an acquaintance, find new friends with similar research interests, or just find someone with whom to share a ride to the airport. This diversity of information needs makes it difficult to generate a static ranked social recommendation list that fits all cases (Chen et al., 2009). While an *optimal* static fusion could provide the best ranking across the cases, it might be *sub-optimal* in each specific case.

The known problems of hybrid recommender systems have encouraged several projects that explored the ideas of controllability and transparency of hybrid recommendation processes (Knijnenburg et al., 2012; Parra and Brusilovsky, 2015; di Sciascio et al., 2018); however, these projects have focused on the *fusion component* and have not addressed the explainability of the contributing sources and their fusion. Our own experience demonstrated that visual interfaces for user-controlled fusion could not assure that these users will understand the underlying rationale of each contributing data or methods; namely, how the recommendation has been made (Jannach et al., 2016). As mentioned above, if a recommendation mechanism is too complicated for non-professional users to comprehend, its transparency could be enhanced by *explainability*; for example, the system may justify the *reasoning process* of why the recommendation was presented (Tsai and Brusilovsky, 2018; di Sciascio et al., 2018). We believe that to increase the transparency of social recommender systems, interactive user interfaces should be augmented with multiple kinds of explanations for each recommendation source or engine. Several interfaces and approaches to provide explanations have been proposed and studied to assess the improvement of user satisfaction and other system aspects (Herlocker et al., 2000; Tintarev and Masthoff, 2012). Explaining recommendations can achieve different *explanatory goals* through either single-style or hybrid explanations (Kouki et al., 2017).

While both controllability and explainability were separately explored as approaches to improve recommendation transparency and user satisfaction, surprisingly little is known about how the user will interact with the system when both the fusion process and reasoning process are transparent. In this paper, we investigate the effects of adding user control and visual explanations to a hybrid social recommender system. We start with the review of relevant work on the topics of interactive recommender systems, as well as the overall controllability and explainability of the recommendation process. We present *Relevance Tuner+*, an extension of our earlier interactive recommender system *Relevance Tuner* (Tsai and Brusilovsky, 2018, 2019c), which provides a controllable and explainable interface for hybrid social recommendations on the *Conference Navigator* platform (Brusilovsky et al., 2016). *Relevance Tuner+* allows the users to control the fusion of multiple recommender sources while also offering explanations of both the fusion process and each of the source recommendations. We also present the design in a controlled study (N=50), which explored the impact of controllability and explainability in the context of recommending attendees to meet at an academic conference. The study brought some interesting results, which are presented and discussed in the remaining part of the paper.

2 Related Works

The rapid growth of both content and data on social applications may cause *information overload*, which makes it harder for users to search for or filter out relevant content (Guy, 2015). Social recommender systems have been widely adopted in many social applications. Various kinds of social data available in social applications, such as social links (Lee and Brusilovsky, 2018), tags (Bogers, 2018), or reviews (O’Mahony and Smyth, 2018) can be used to rank relevant items through ranking-based collaborative filtering (Kluver et al., 2018). Social recommender systems can also be used to reduce *social overload*, such as making people-to-people connections. Guy et al. (2009) proposed a *people you may know* recommendation function to follow new friends on a corporation-based social media platform. Guy (2018) provided social recommendation for the user to follow interesting people. These works can enable the users to establish new personal relationships or even in professional connections, such as a research collaborator in an academic context (Sie et al., 2012).

Bringing user control to a hybrid recommender system allows the users to have an immediate effect on their recommendations (Jannach et al., 2016); namely, the users can further filter or re-sort recommendations based on their preference or information needs. This usually requires an interactive visualization framework that combines recommendation with visualization techniques to support user interaction or intervention into the recommendation process (He et al., 2016). The idea of a user-controllable interface of different recommendation approaches was originally presented in the work of Schafer et al. (2002). Bostandjiev et al. (2012) suggested a slider-based interface that allows the user to adjust the weights of the items and the social connections. Verbert et al. (2013) encouraged users to choose the most appropriate sources of relevance for each case and provided a cluster-map interface to support user-driven exploration and control of tags, agents, and users. Parra and Brusilovsky (2015) attempted to increase both controllability and

transparency of hybrid recommendations by using a combination of sliders for controlling the fusion and a Venn diagram to visualize results. Ekstrand et al. (2015) discussed a recommender-switching feature to let the users choose recommender algorithms. Tsai and Brusilovsky (2018) added user-controllable interfaces, a two-dimensional scatter-plot, and multiple relevance sliders to a social recommender system for conference attendees. Bailey et al. (2018) further provides visualization for data analytic task using the conference data.

User controllability has also been recognized as a crucial component in supporting exploratory search, namely allowing users to narrow down the number of items and inspect the details during the information-seeking process (di Sciascio et al., 2018). Ahn et al. (2010) presents a summary of search results in the form of entity clouds, which allows users to explore the results in a controllable interface. Han et al. (2013) offered users an option to re-sort people search results based on multiple user-related factors. di Sciascio et al. (2016) proposed an interface for understanding, refining, and reorganizing documents. di Sciascio et al. (2018) integrated controllable social search functionality into an exploratory search system. An effective interactive visualization representation can enable users to control the process of recommendation (He et al., 2016). The users have a different form of control over the recommendation process: 1) let the user tell the system what they like through forming a user profile (Jannach et al., 2016) or through an adaptive dialog (Kobsa et al., 2001); 2) allow the user to adjust the recommendation profile (Harper et al., 2015) to fuse recommendations from different sources of relevance (Tsai and Brusilovsky, 2018; di Sciascio et al., 2018) or to influence the presented layout (Bostandjiev et al., 2012; Verbert et al., 2013; Parra and Brusilovsky, 2015).

An alternate approach to increase the recommendation process transparency and user satisfaction explored in the literature is providing explanations for recommendations (Chen and Wang, 2017). Explanations that expose the reasoning behind a recommendation could especially increase system transparency (Tintarev and Masthoff, 2015). Many scholars have suggested different explanation functions to increase the inspectability of the recommender system (Kouki et al., 2017). The function provides the transparency that let users realize how the system works (Tintarev and Masthoff, 2012; Herlocker et al., 2000). The exposure of the recommendation process through visual interfaces can also increase the inspectability of the system (Knijnenburg et al., 2012). Many different types of research have been done on this subject. For example, Tsai and Brusilovsky (2017) provides recommendation visualization to increase the transparency of the recommender system. Verbert et al. (2013) provides a set-based visualization to let the user explore the desired recommendation items. Other researchers further indicated that the value of explaining interfaces could enhance user experiences. The explanation interface was associated with the perception of recommendation quality (Tintarev and Masthoff, 2012), gaining trust in the system (Cramer et al., 2008), and experiencing the competence of the system (Wang and Benbasat, 2007). The studies of Gretarsson et al. (2010); O'Donovan et al. (2008, 2009) have all mentioned that providing a controllable interface in the social recommender system can increase overall user satisfaction. These authors adopted interactive graphical interfaces to present the social recommendations that enable controllability of an item or user-level preference in a collaborative recommender system.

Enhancing explainability in recommender systems has drawn more and more attention. Researchers in the field of recommender systems explored a range of

advanced interfaces to support exploration, transparency, explainability, and controllability of recommendations (He et al., 2016). Explaining recommendations can achieve different *explanatory goals* by single-style or hybrid explanations (Tintarev and Masthoff, 2012; Sharma and Cosley, 2013; Kouki et al., 2017). A number of explanation interfaces and approaches have been proposed and studied to assess the improvement of user satisfaction and other aspects (Tintarev and Masthoff, 2012). However, most of the evaluation has focused on solely user perception or preferences (Papadimitriou et al., 2012; Kouki et al., 2017). In most cases, it remains unclear whether different kinds of explanations could improve the objective parameters of user performance rather than their perception of which option is better while *inspecting* the explanation interface (Knijnenburg et al., 2012).

The *Controllability* and *Explainability* contribute to the recommender system transparency in different aspects. *Controllability* enabled end users to participate in the recommendation process by providing various kinds of input (Bostandjiev et al., 2012; Parra and Brusilovsky, 2015; Tsai and Brusilovsky, 2017), such as adjusting preferences or exploring recommendations. *Explainability* features allowed interactive recommender systems to deal with the “black-box” problem, i.e., to explain the inner logic of the recommendation process to the end users (Swearingen and Sinha, 2001; Vig et al., 2009). A visual interface for the user-controlled hybrid fusion of recommender sources cannot assure that the users understand the underlying rationale of each contributing recommender; namely, the recommendation algorithm (Jannach et al., 2016). In the case when recommendation mechanism is too complicated to non-professional users to explain, some considerable transparency could be achieved by *explainability*; i.e., the system may just need to *justify* why the recommendation was presented through user interface designs (Tintarev and Masthoff, 2011; Tsai and Brusilovsky, 2018; di Sciascio et al., 2018) or natural language explanations (Musto et al., 2019a,b).

We conducted several preliminary user studies to prepare the experiment platform of this paper. We first focused on finding an effective design for user-controllable interfaces, which allow users to fuse and control multiple relevances for influencing the presentation of recommendations Tsai and Brusilovsky (2017, 2018). We further followed a participatory design approach to summarize the key components of each recommendation model and then build explainable user interfaces based on user preference and performance Tsai and Brusilovsky (2019a,b). We further deployed the integrated experiment platform (i.e., the controllable and explainable user interfaces) to a small-scale online user study Tsai and Brusilovsky (2019c). Based on the leaned experience and user feedback, the novelty of this paper is to combine the proven effective controllable and explainable user interfaces and to evaluate the design with a large-scale user experiment.

3 Controllable and Explainable Social Recommender Interfaces

3.1 User Interface Design

To explore the value of controllability and explainability in the context of recommenders, we developed a social recommender interface for attendees of academic conferences. The system, called *Relevance Tuner+*, was implemented as a component of a conference support system - Conference Navigator (Brusilovsky

et al., 2016). It is an extension of our early controllable recommender system, RelevanceTuner, with several explanation-focused features (Tsai and Brusilovsky, 2018, 2019c). The original RelevanceTuner system allowed users to “tune” (re-rank) the list of recommended co-attendees to meet using four sliders (see Fig. 1, Section A). Each slider controlled the importance of one of the four source recommender engines of the hybrid recommender engine - *publication similarity*, *topic similarity*, *co-authorship similarity*, and *interest similarity*. By moving the slider to the right, the relative importance of the component increases, and by moving it to the left, the importance decreases. The final impact of a specific source on the ranking of a specific co-attende was determined by both the relevance of this attendee along this source (for example, the similarity of her publication to the publications of the target user) and the current importance of the source selected by the user (e.g., the importance of the Publication Similarity source). If the slider is set to the minimum value (importance = 0), the recommender engine that did not contribute to the ranking will be disabled. If the slider is set to the maximum value (importance = 10), then the recommender engine’s importance will be doubled.

To make this hybrid fusion process more transparent, the RelevanceTuner interface uses a colored stackable bar visualization to show how the total relevance score of a recommended attendee was composed of source-level relevance scores (see Fig. 1, Section B). More information about recommended people could be obtained by using the links in Section D. The RelevanceTuner+ further expanded this controllable interface with a *explanation icon* shown next to each recommendation in Section C, which could be used to *inspect* the relevance of this recommendation. Following a click on the explanation icon, a window will pop up (see Fig 2) to show a clickable explanation table, presenting visual explanations of calculated relevance for the selected recommendation engine.

The pop-up explanation window provides access to four explanation interfaces through color-coordinated buttons. Each interface focuses on explaining the recommendation score delivered by the corresponding component of the hybrid engine: a) two-way bar chart (for *Publication similarity*) showed the mutual relationship between two scholars publication terms and term frequency; b) topical radar (for *Topic Similarity*) presented the top topics of the user and compared them with the selected attendee; c) strength graph (for *Co-Authorship similarity*) explained the strength of social connections by presenting the co-authorship network connecting the recommended attendee to the target user (nodes and edges are representing authors and co-authorship, respectively); d) Venn tags (for *CN3 Interest similarity*) visualized the similarity of co-bookmarked, or co-followed item in the form of a Venn diagram. The design can be found in Fig. 3. The specific visualization was constructed through a participatory design process, presented in our previous study (Tsai and Brusilovsky, 2019a).

The *Relevance Tuner+* enhanced recommender system transparency in three ways. First, the user was able to inspect the static colored stackable bar to understand the fused relevance score in a multi-relevance context. Second, the user was able to influence the fusion by changing the weights and observing the immediate changes of the colored relevance bars and item ranking. This explorability makes the fusion process transparent. Third, the user can access more information about the reasoning process by clicking the explanation icons. In the design of *Relevance Tuner+*, we attempted to separate the transparency of the fusion

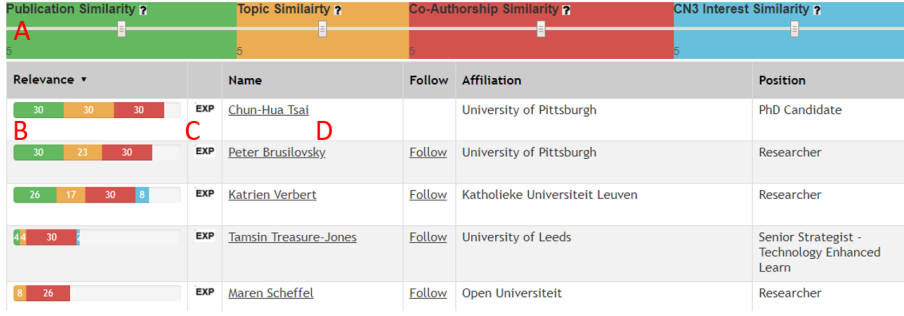


Fig. 1 The design of Relevance Tuner+: (A) relevance sliders; (B) stackable score bar; (C) explanation icon; (D) user profiles.

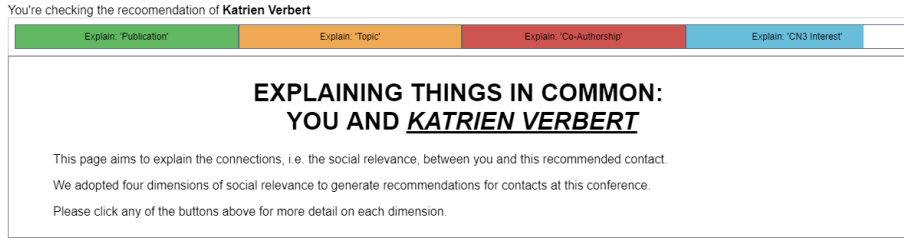


Fig. 2 The pop-up window of clicking the *explanation icon* (shown in Figure 1 Section C).

process from the explanation of each level of relevance obtained by clicking the explanation icon. To be more specific, we intend to improve system transparency in the recommendation process, as well as in the reasoning process.

To ensure the soundness of RelevanceTuner+, all its components were separately evaluated in a sequence of studies. The evaluation of the controllable slider was presented in our previous study (Tsai and Brusilovsky, 2018). The findings showed that the visual interfaces significantly reduced the information search efforts tasks and helped users to more accurately perceive recommendation quality, as well as an improvement in overall user satisfaction. The design of the explanation icon was discussed in our previous studies (Tsai and Brusilovsky, 2019c); the findings indicated the effectiveness of the proposed explanation models and a significant improvement in the perception of explainability, but we found that providing controllability and explainability was complementary in that the users may not adopt both of the functions together. The visual interfaces for the four source explanation models were selected through a stage-based participatory design process to interactively discuss the user preference of the interface prototypes (Tsai and Brusilovsky, 2019a). We further improved the design with a performance-focused evaluation. The result suggests that the user-preferred interface may not guarantee the same level of performance. To resolve the disagreements between designs favored by user preferences and efficiency evaluation, we conducted a user study (Tsai and Brusilovsky, 2019b) to determine the best designs for the recommendation models where the choice of the top design was not evident after the two earlier studies. The following section reviews the final version of the source recommendation model visualizations in more detail.

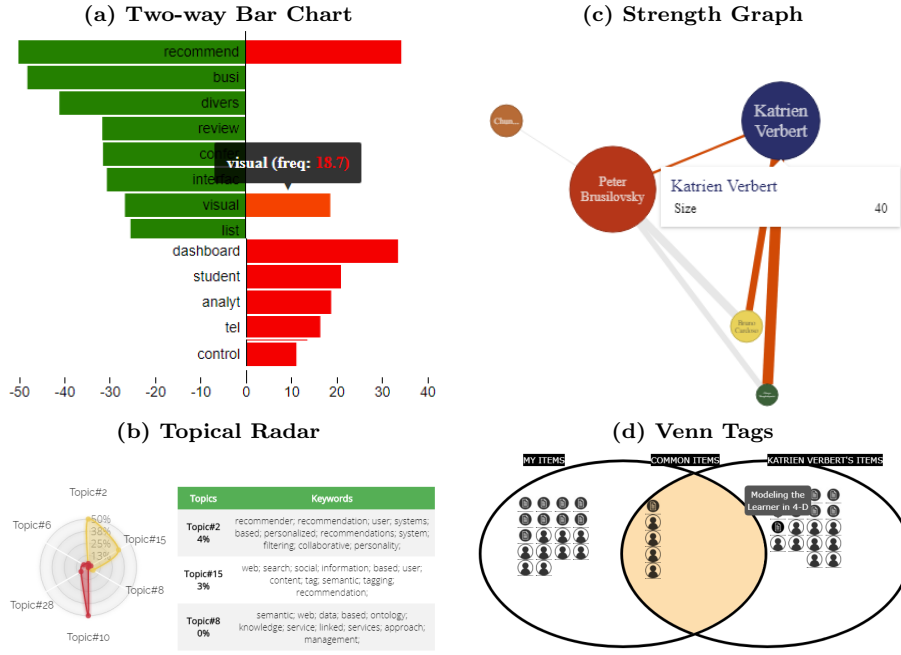


Fig. 3 Explanation of interfaces: (a) Two-Way bar chart for publication similarity; (b) Topical radar for topic similarity; (c) Enhanced strength network for co-authorship network; (d) Venn tag for CN3 interest similarity.

3.2 Explaining Recommendation Models

A total of four recommender models were introduced in RelevanceTuner+: 1) *Publication similarity*: cosine similarity of users' publication text; 2) *Topic similarity*: topic modeling similarity of research interests (topics); 3) *Co-authorship similarity*: the degree of network distance, based on a shared co-authorship network; and 4) *CN3 interest similarity*: the number of papers co-bookmarked, as well as the authors co-followed. The detailed design of the explanations can be found here (Tsai and Brusilovsky, 2019a).

3.2.1 Publication similarity:

This similarity was determined by the degree of text similarity between two scholars' publication vectors using cosine similarity. We applied tf-idf to create the vector with a word frequency upper bound of 0.5 and a lower bound of 0.01 to eliminate both common and rarely used words. In this model, the key components were the *terms* of the paper title and abstract, as well as its *term frequency*.

On the basis of our studies (Tsai and Brusilovsky, 2019a,b) we adopted a *two-way bar chart* visualization as an approach to explaining the text-level similarity between the publications of the user and the attendee (Fig 3a). The visualization presented the mutual relationship of two scholars' publication terms and term frequency; i.e., one scholar on a positive scale and the other scholar on a negative

scale. This visualization presented the *terms* of the paper title and abstract. The bar length indicates the term frequency in the documents. The user’s terms and the attendee’s terms are presented on the left and right, respectively. The user can inspect the *words-in-common* through the term appear in both sides, e.g., the term *visual* in Fig. 3a means the term appeared in both of the scholars’ publications. We ran a crowdsourcing study to determine the setting of the bar chart. Based on the study result, we chose 30 terms (versus 60 terms) ordered by individual relevance (versus the sum of relevance).

3.2.2 Topic similarity:

This similarity was determined by matching research interests using topic modeling. We used a latent Dirichlet allocation (LDA) to attribute collected terms from publications to one of the topics. We chose 30 topics to build the topic model for all scholars. Based on the model, we then calculated the topic similarity between any two scholars. The key components were the *research topics* and the *topical words* of each research topic (Wu and Ester, 2015).

We presented research topics in a radar chart and the topical words of each research topic in table (Wu and Ester, 2015). The visualization design can be found in Fig. 3b. The radar chart was presented on the left side. We selected the top 5 (ranked by *beta* value from a total of 30 topics) topics of the user and compared them with those of the other scholar. A table with topical words was presented on the right so that the user can inspect the context of each research topic. We found this design to be effective based on the user study of (Tsai and Brusilovsky, 2019c). Based on the study result, the users were able to achieve 97% correct rate of sorting multiple recommendation models, solely using the visualization.

3.2.3 Co-authorship similarity:

This similarity approximated the co-authorship network distance between the source and recommended users. For each pair of scholars, we tried to find six possible paths for connecting them, based on their co-authorship relationships. The network distance is determined by the average distance of the six paths. The path was defined as a single shortest path between the source vertex (the source user) to the target vertices (the recommended users). We used the breadth-first search for unweighted graphs and Dijkstra’s algorithm for weighted graphs. For finding six paths, we repeated the search six times, the *found* shortest paths would be removed from the graph to avoid the same path.

We presented a co-authorship network in a path graph (Tsai and Brusilovsky, 2017). The key components were the *coauthors* (as nodes), *coauthorship* (as edges) and the *distance of connection between the two scholars*. Fig. 3c shows the visualization design. For connecting the user (yellow circle on the left) to the attendee (red color in the right), we tried to find six possible paths (one shortest and five alternatives) by direct and indirect co-authorship. Our original attempt to design a visual explanation for co-authorship similarity (Tsai and Brusilovsky, 2019b) was not successful; we found that users failed to use the information presented in this visualization in sorting the recommendations. The current design was selected through an additional crowdsourced study. The study was a 2x2 factor design that has four conditions: edge thickness as the relevance between the connected nodes

(i.e., the co-authored papers between two scholars) and node size as the number of papers (i.e., to make the node size represent the information about the additional number of papers). Based on the results of the study, we decided to set 1) node size as the number of papers; 2) edge thickness as the number of co-authored papers. The improved design has been shown to be effective in obtaining a 60% correct rate of sorting recommendations by relevance.

3.2.4 Interest similarity:

This similarity was determined by the number of co-bookmarked conference papers and co-connected authors in the conference support social system Conference Navigator (CN3). We used the number of shared items as the CN3 interest similarity. The key component is the shared *conference papers* and authors.

We presented co-bookmarked papers in a design of *Venn tags* (shown in Fig. 3d). The study of (Kouki et al., 2017; Parra and Brusilovsky, 2015) has pointed out the user preferred the Venn diagram as an explanation in a recommender system. The interface shown in Fig 3d: Venn Tags, we implemented the same idea with the bookmarked items. The idea is to present the bookmarked item, using an icon, in the Venn diagram. The two sides are the bookmarked item belong to one party. The co-bookmarked or co-followed item will be placed in the middle. The users can hover over the icon for detailed information, such as the paper's title or author's name. We found this design to be effective based on the user study of (Tsai and Brusilovsky, 2019c). Based on the study result, the users were able to achieve 93% correct rate of sorting multiple recommendation models by solely using the visualization.

4 Research Questions

The integration of controllability and explainability in a single interactive recommender system allowed us to address two important research questions.

[RQ1] How do controllability and explainability affect the *user perception*, *user experience*, and *user engagement* with a hybrid social recommender system?

[RQ2] Is there an interaction effect between controllability and explainability in a hybrid social recommender system?

Many studies have investigated the effects of controllable (O'Donovan et al., 2008; Parra and Brusilovsky, 2015; He et al., 2016; Ekstrand et al., 2015) or explainable user interfaces (Herlocker et al., 2000; Tintarev and Masthoff, 2015; Bostandjiev et al., 2012; Verbert et al., 2013; Parra and Brusilovsky, 2015). However, neither a direct comparison of these effects nor the interaction effect between controllability and explainability was examined. To answer these research questions, we designed and performed a controlled human subject study presented in the following sections of the paper. In order to prepare for this study and to assure the effectiveness of the interfaces, we conducted several preliminary user studies (Tsai and Brusilovsky, 2018, 2019a,c,b). These pre-studies focused on finding an effective design of user-controllable interfaces, which allows users to fuse and control multiple relevances for influencing the presentation of recommendation. We further followed a participatory design to summarize the key components of each

	UMAP 2015	UMAP 2016	UMAP 2017	UMAP 2018
Number of Papers	143	129	168	108
Number of Authors	231	305	345	289
Number of Attendees	116	115	151	131
Number of Bookmarks	664	660	714	342
Assigned Interface	BASE	CONT	EXPL	FULL

Table 1 Meta-data of the UMAP dataset

recommendation model, and then built explainable user interfaces based on user preference and performance.

5 User Study

In our study, we attempted to explore the effects of controllability and explainability in a realistic recommendation context. This goal was facilitated by the practical nature of RelevanceExplorer+, which was implemented and used as a social recommendation component or a popular conference support system Conference Navigator 3 (CN3), which has been used at many research conferences (Brusilovsky et al., 2016). However, the real conference context makes it impossible to run a reliable, controlled study. While in our past work, we did explore earlier versions of RelevanceExplorer in conference-based field studies (Tsai and Brusilovsky, 2018), the needs of a randomized controlled study caused us to perform this study outside of a real conference context. However, we used data from real conferences and engaged experienced graduate students who were a close approximation to real conference participants. The study used different scenario-based information tasks and data-driven analysis methods.

5.1 Data and Participants

The recommendations produced by all four recommendation models are based on data collected by the CN3 system. We used real data from UMAP Conferences from 2015 to 2018. UMAP is the premier international conference for researchers working on *personalized recommender systems, technology-enhanced adaptive learning, personalized social web adaptive hypermedia* and *the semantic Web*. The dataset contained conference proceeding data, including conference papers (author, title, and abstract), author and attendee list (name, published papers, affiliation, and position) and user feedback (bookmarks on papers and authors). As in the real conference context, our study used recommendation models to provide a social recommendation, i.e., recommending conference attendees to meet. The summary description of the UMAP dataset can be found in Table 1. Subjects were recruited by emails and ads posted at the building of the School of Computing and Information at the University of Pittsburgh. The promotional emails were sent to mailing lists of all doctoral students and summer registered masters' students. The main requirement was that the subjects should be able to perform an information search on web application as well as had a declared major in Library Science (LIS), Information Sciences (IS), Telecommunications (MST), Computer Sciences (CS) and the Intelligent Systems Program (ISP) at the School

of Computing and Information. Each participant received USD\$20 compensation and signed an informed consent form.

We conducted a power analysis to determine the overall size of participants. A calculation of power analysis involves the following four parameters. 1) Alpha (α): a cut-off p-value that indicates threshold probability for rejecting the null hypothesis (Type I error rate), the $\alpha = 0.05$. 2) Power: the probability of finding a true effect - the probability of failing to reject the null hypothesis under the alternative hypothesis (Type II error rate). We choose the power value as 0.8. 3) Effect size: the expected effect size, which refers to the expected correlation coefficient in this case. The goal is to find medium-sized effects where the value is 0.35. 4) Sample size N: this value determines the participant's size to maintain statistical power. We used *G*Power* software to calculate the sample size, which results in the calculation result of a total sample size of 49. The sample size supports the actual statistical power of 80%.

Based on the power analysis, a total of 50 participants (N=50) were recruited for the user study from May 30 to June 16, 2019, at the University of Pittsburgh. The subjects were 28 men and 22 women whose age ranged from 22 to 44 (M=28.82, SD=4.83; M=Mean, SD=Standard Deviation). A total of 22 masters' students joined the study, including 21 IS and 1 MST majors. There were 28 doctoral students, including 18 IS, 3 LIS, 2 CS, and 5 ISP majors. All doctoral students had at least one publication and one conference attending experience, but no Masters' students had any publication or any conference attendance. Subjects took between 52 and 192 minutes (M=106.05, SD=28.80) to complete the study.

5.2 Experiment Design and Procedure

In this study, we explored *RelevanceTuner+*, a controllable and explainable user interface for social recommendation. To access the separate and combined value of its controllability and explainability features, we used a 2x2 experimental design with four conditions. There were two interventions in this design: enable/disable relevance sliders (Section A in Fig 1) and enable/disable explanation icon (Section C in Fig. 1). The four conditions produced by these interventions are: Baseline interface (**BASE**), with both slider and explanation icon disabled; Controllable interface (**CONT**), with slider enabled; Explainable interface (**EXPL**), with explanation icon enabled; and Full interface (**FULL**), with both slider and explanation icon enabled. Each condition was used with a specific year of UMAP conference data (see Table 1). We followed within-subject design, so all participants were asked to use each interface for one training session and two study tasks, and to fill out a post-stage questionnaire at the end of their work with each interface. At the end of the study, participants were asked to fill out a post-study questionnaire. To minimize the learning effect and bias, we followed a Latin square design to balance the conditions given to each participant. The study procedure can be summarized as follows:

1. Pre-study questionnaire (for user preference elicitation)
2. Interface 1: training session, task 1, task 2
3. Post-stage questionnaire
4. Interface 2: training session, task 1, task 2

5. Post-stage questionnaire
6. Interface 3: training session, task 1, task 2
7. Post-stage questionnaire
8. Interface 4: training session, task 1, task 2
9. Post-stage questionnaire
10. Post-study questionnaire

User Preference Elicitation: In the real-world conference scenario, most of the event attendees have some publications that the system can use to generate social recommendations. However, in our study, only some of the subjects have ever published papers. To better control the data sparsity, we asked the participants to complete a pre-study questionnaire before the user study. The pre-study questionnaire sought to conduct the user preference elicitation process so we can generate personalized social recommendations. The pre-study questionnaire had four questions. First, we asked subjects to pick five preferred research topics (out of 30 topics) that identified by the *Topic Similarity*. Each topic came with hundreds of *topical keywords*. We listed the top 10 keywords to let the user understand the context of each topics. For example, “*Topic#1: behavior, user, cultural, networks, social, world, web, differences, mixed, models, effect*”. We used the selected topics for matching *topic similarity*. The topical keywords of the chosen topic were then used to calculate *publication similarity*. Second, we asked subjects to select five *mentors* from the faculty directory of the School of Computing and Information. We filtered 18 professors (out of 41 tenure-track professors) whose research interests were close to the topics of the UMAP conference. The selected mentors will be treated as the subjects’ *coauthors* for calculating co-authorship similarity. Third, we asked subjects to select 3 favorite papers from each year’s UMAP top 10 most bookmarked papers (a total of 12 papers). The data was used to calculate *CN3 Interest Similarity*.

In the study, the subjects were told to act as a researcher who is attending the conference. They were requested to pick a suitable candidate among conference attendees, based on their best judgment. Participants were given one training session and two information search tasks for each interface. In the training session, we urged the user study participants to follow a few steps, so they have a chance to familiarize themselves with the system. The instrumentation details are listed below.

Training Session:

1. Inspect the definition of “publication similarity” (click on the question mark next to it) and read the text that appears.
2. Re-tune one or more sliders and inspect how adjusting that changes the stackable color bar.
3. Click two scholar names and check their publications.
4. Click two “Explanation Icon” and inspect the four explanation functions in them.
5. Follow any four scholars by clicking the Follow button; you will need to provide a reason for your selection. Please type “test” as your answer.

To be noted, in step 1, the inspecting similarity rotated across conditions to cover all four similarities. Step 2 only appeared in the *CONT* and *FULL* interfaces. Step 4 only appeared in the *EXPL* and the *FULL* interface. In step 5, a pop-up box

appeared to collect user feedback/comments when the user clicked on the “follow” button (see Fig. 1, “Follow” column).

Study tasks 1 and 2 were explicitly designed as relevance-oriented and diverse-oriented information search tasks, respectively. We tried to design the assigned tasks as realistic tasks that could be naturally pursued by attendees at research conferences. For the relevance-oriented task (Task 1), we expect to see the user to coordinate multiple relevance aspects and efficiently find out the desired candidates. In contrast, in the diverse-oriented task (Task 2), we expect the user to select candidates from different relevance aspects, which provides a diverse selection. Both of the tasks asked the user to “follow” four scholars from the conference attendee list, based on different criteria. To promote the function usage and advance the information need, in each “following”, we asked the user to “justify” their choice by typing a couple of sentences about the reasons they decided to follow the chosen scholar (just as in Step 5 above).

Task 1: Find Advisor/Mentor

- If you plan to pursue a doctoral degree after your current degree program, you now have an excellent opportunity to find a prospective advisor or mentor at this conference. For this task, you will select scholars to follow as potential advisors/mentors.
- Please “follow” four scholars whose work is more relevant to your research interest(s). The ideal candidates will be scholars who the system identified as more connected to your chosen SCI professors (so they can provide you a strong recommendation letter).
- You are also expected to justify your selections (for example, to the Ph.D. admission committee), so it is important to pay attention to why you make your selection.

Task 2: Find a Guest Speaker

- You have been asked to invite guest speakers for an academic seminar at your home school. The main seminar theme is to encourage the inclusion of different types of research and community.
- Please “follow” four scholars from the conference whose work matches the seminar theme. The ideal candidates are: 1) those who can represent the inclusion of different types of research, i.e., those scholars are expected to work on different research topics; 2) those who are “less” connected to your chosen SCI professors.
- You are also expected to justify your selections (for example, to the dean’s office), so it is important to pay attention to why you make your selection.

5.3 Measurements

We used both subjective metrics and objective metrics to measure the effectiveness of the controllable and explainable interfaces. The subjective measures were captured by questionnaires. We used the existing constructs (groups of questions) from the works of (Pu et al., 2011; Knijnenburg et al., 2012; Knijnenburg and Willemsen, 2015). In this study, we collected user feedback in seven constructs:

Construct	Factor		Score M (SD)
General Trust in Technology (PC1)	1	I feel technology never works.	1.34 (0.65)
	2	I'm less confident when I use technology.	1.76 (0.93)
	3	The usefulness of technology is highly overrated.	2.82 (1.49)
	4	Technology may cause harm to people.	3.52 (1.95)
	Average		2.36 (1.59)
User Characteristics: Scholarly Expertise (PC2)	1	Compared to my peers, I have a lot of collaborators or research experiences.	3.84 (1.62)
	2	Compared to my peers, I am an expert on the subject of the conference.	3.40 (1.51)
	3	I only know a few scholars at the conference.	5.04 (1.51)
	4	I frequently attend academic conferences.	3.50 (1.72)
	Average		3.94 (1.71)
General Acceptance of Diversity (PC3)	1	I'd like to see scholars with dissimilar research interests.	6.04 (1.02)
	2	I would be satisfied if I was recommended unfamiliar items.	4.54 (1.63)
	3	I think the recommendation should cover the scholars from different research areas.	5.96 (1.41)
	4	I like to see the recommendations beyond my interests.	5.48 (1.32)
	Average		5.05 (1.48)
System-specific Privacy Concern (SC1)	1	I'm afraid the system discloses private information about me.	3.26 (1.84)
	2	Personalized recommender usually invades my privacy.	5.56 (1.83)
	3	I'm uncomfortable providing private data even if it helps me to receive better recommendations	3.30 (1.75)
	4	I think the recommender system should respect the confidentiality of my data.	6.00 (1.37)
	Average		4.03 (2.04)
System-specific Familiarity and Understanding (SC2)	1	Compared to my peers, I am familiar with the technology of recommender systems.	4.80 (1.47)
	2	I feel comfortable to solve some mathematical or equation questions.	5.74 (1.32)
	3	I am confident when I first time interacts with a new information system.	5.26 (1.33)
	4	I am familiar with programming language.	6.02 (1.31)
	Average		5.45 (1.43)

Table 2 The result of post-study survey: we collected three constructs of personal characteristics (PC) and two constructs of system-specific characteristics (SC). We used a 7 Likert scale in the survey.

perceived recommendation quality/accuracy(Q); perceived recommendation diversity/variety(D); perceived control (C); perceived system effectiveness (E); perceived trust (T); perceived transparency (P); and satisfaction (S). All these constructs were collected in the stage of the post-stage survey after the participants interacted with each experiment condition. The questions are listed below, and the survey results were shown in Table 4.

- Construct Q: Perceived Recommendation Quality/Accuracy
 - Q1: The recommender provided good recommendations.
 - Q2: I liked the recommendations provided by the system.
 - Q3: The recommended scholars fitted my preference.

- Q4: I did not like any of the recommended scholars.
- Construct D: Perceived Recommendation Diversity/Variety
 - D1: The recommender helped me discover new contacts at the conference.
 - D2: The scholars that were recommended to me are diverse.
 - D3: The list of recommendations included scholars from many different research areas.
 - D4: The list of recommendations was very similar.
- Construct C: Perceived Control
 - C1: The recommender allowed me to modify my preference.
 - C2: I became familiar with the system very quickly.
 - C3: The layout of the recommender interface was adequate.
 - C4: The recommender helped me to make the following decisions faster.
- Construct E: Perceived System Effectiveness
 - E1: Using the system is a pleasant experience.
 - E2: I made better choices with the recommender.
 - E3: I found better items using the recommender.
 - E4: I felt bored when using the recommender.
- Construct T: Perceived Trust
 - T1: I am convinced by the scholar recommended to me.
 - T2: I am confident I will like the items recommended to me.
 - T3: The recommender made me more confident about my selection/decision.
 - T4: The recommender can be trusted.
- Construct P: Perceived Transparency
 - P1: The provided information was sufficient for me to make a good decision.
 - P2: The recommender explained why the scholars were recommended to me.
 - P3: I understood why the scholars were recommended to me.
- Construct S: Satisfaction
 - S1: I will use this recommender again.
 - S2: I will tell my friends about this recommender.
 - S3: Overall, I am satisfied with the recommender.
 - S4: The recommender helped me find the ideal contacts at the conference.

We further collected three constructs of personal characteristics (PC) and two constructs of system-specific characteristics (SC) in the post-study survey. These constructs included general trust in technology (PC1), scholarly expertise (PC2), general acceptance of diversity (PC3), system-specific privacy concern (SC1), and system-specific familiarity and understanding (SC2). The constructs help to understand the user personality and expertise that were positively correlated with user perceptions (Schaffer et al., 2019). The detailed questions and survey results from both the PC and SC constructs can be found in Table 2.

We used objective metrics to measure user engagement and recommendation quality. In all four study interfaces, the user can *inspect* the recommendation by clicking on different links and interface elements. We logged five user actions in this user study, including clicks on *Learn more*, *Scholar Name*, *Relevance Tuner*, *Explanation Icon* and *Explanation Tabs*. We also logged the *time spent* to measure the time used in each task. In addition to user engagement measures, we adopted four metrics to evaluate the recommendation quality (Parra and Brusilovsky, 2015): 1) *Top N*: this measured the accuracy of a list of top k recommendation. 2) *Mean*

Actions		BASE (S1)		CONT (S2)		EXPL (S3)		FULL (S4)	
		M(SD)	UC	M(SD)	UC	M(SDE)	UC	M(SD)	UC
T	Learn More	0.72 (1.52)	15	0.74 (1.58)	14	0.52 (0.97)	16	0.36 (0.98)	9
A	Scholar Name	13.66 (10.80)	44	15.12 (16.03)	43	10.52 (14.24)	41	11.72 (17.10)	40
S									
K	Tuners	-	-	30.26 (39.97)	42	-	-	22.38 (25.67)	42
1	Exp. Icon	-	-	-	-	7.28 (8.66)	39	7.42 (10.74)	35
	Exp. Tabs	-	-	-	-	18.56 (23.63)	37	15.22 (18.44)	34
	Time Spent (in Seconds)	517.24 (343.80)	50	633.46 (682.92)	50	613.84 (409.65)	50	533.32 (303.71)	50
T	Learn More	0.24 (0.82)	6	0.22 (0.54)	8	0.26 (0.87)	6	0.50 (1.24)	9
A	Scholar Name	14.36 (14.35)	44	14.42 (11.89)	45	9.92 (9.86)	43	14.26 (16.79)	41
S									
K	Tuners	-	-	19.34 (21.84)	44	-	-	20.22 (25.81)	40
2	Exp. Icon	-	-	-	-	5.56 (7.70)	29	5.64 (8.26)	32
	Exp. Tabs	-	-	-	-	11.26 (12.88)	29	12.88 (21.98)	31
	Time Spent (in Seconds)	418.28 (310.92)	50	479.24 (341.98)	50	460.12 (289.90)	50	481.56 (299.24)	50

Table 3 User action summary: the table shows the user interaction statistics while performing each of the two tasks using four interfaces. The interactions are defined as: “Learn More”=number of clicks on the question marks (Figure 1, Section A); “Scholar Name”=number of clicks on the name links (Figure 1, Section D); “Tuners”=number of clicks on the relevance tuners (Figure 1, Section A); “Exp. Icon”=number of clicks on the explanation icon (Figure 1, Section C); “Exp. Tabs”=number of clicks on the explanation tabs (top of Figure 2); “Time Spent”=time spent in each interface. (Abbreviation: UC=User Count; M(SD)=Mean(Standard Deviation))

reciprocal rank (MRR): MRR was determined by the ranking position of the first relevant element that was matched (Baeza-Yates et al., 1999). 3) *Normalized discounted cumulative gain (NDCG)*: This metric penalized the lower-ranked relevant items, which approached the relevant item to be ranked in the top positions (Manning et al., 2010). 4) *alpha-NDCG*: this metric is relevant to NDCG but was used to measure for diversified search, where it is appreciated by the number of covered intents (Clarke et al., 2008). All these objective metrics provide a sketch and description of the user interaction with the proposed interfaces.

6 Results

6.1 Action Analysis

Table 3 presents system usage (number of clicks and time spent) of two study tasks in four different conditions. Most importantly, the data indicates that the participants extensively used both control tuners and the explanation icon when these options were provided. There were 42 (out of 50) users who used the tuners to solve the tasks and 39 (out of 50) users who used the explanation icon when solving

the tasks. The users usually inspected 2-3 explanation tabs each time after clicking the explanation icon. We found that the users used the tuners and explanation icon more intensively in the first task than the second task. This is consistent with our previous finding (Tsai and Brusilovsky, 2018) that the user engagement rate is higher when the user interacts with an intelligent interface for the first time. The time spent supports the argument that it took less time to solve the second tasks. The analysis indicates that the need to examine details about a specific candidate user by clicking *Scholar Name* is noticeably decreased when explanations are available; i.e., when a user clicks *Scholar Name* less frequently in the *EXPL* interface and more frequently in the *CONT* interface. It hints that the explanations do help to understand the match. However, we did not find significance for *Learn More*, *Scholar Name*, and *Time Spent* metrics when we tested the user interaction parameters using a two-way ANOVA analysis.

We further analyzed the user interaction with the control and explanation functions. Fig. 4 presents the average re-weighting value of *CONT* and *FULL* interfaces through four box-plots. The average re-weighting value was calculated by $\frac{\sum_{i=1}^n Slider_s(n)}{n}$, where n represents the number of clicks on the slider s and $S_s(n)$ represents the selected weight for each slider adjustment attempt (from 0 to 10). The usage analysis shows that the users tend to increase the relevance weights from its default weighting value of 5 in relevance-oriented Task 1, when no compromise between sources is necessary, and decrease the weights in diversity-oriented Task 2, when a balance between sources has to be found. We can observe a similar pattern across all four relevance sliders and both of the interfaces. In other words, the users tend to *enhance* the source importance when seeking similar items, but *reduce* the source importance when seeking dissimilar items. The effect was most visible in adjusting publication and topic similarity, which were primary targets in both tasks, and was least visible for co-authorship similarity in Task 1, in which it was a secondary (although positive) consideration. However, we found the effect to be less obvious in the group of co-authorship similarity. This data hints that the users understood the purpose of the sliders and used them in accordance with the nature of the assigned information tasks.

6.2 Recommendation Quality

Given the well-known tendency of search and recommender system users to focus on the top positions in the ranked list, a system that helps the user find relevant people high in the ranked list offers a better recommendation quality than the one that pushes the user to go to the lower ranks. To measure recommendation quality, we treated user selections as a ground truth, assuming that the graduate-level users are the best judges of who is the right person for them to meet. For example, in the “Find Advisor/Mentor” scenario (Task 1), we used the user-selected scholars (recommendation items) as ground truth. The users may be able to re-rank the recommendations in the controllable interfaces, but we simply focused on the ranking they selected (the frequency of re-ranks were ignored). In essence, all metrics used in our evaluation attempted to measure in different ways to what extent a specific version of the system, either with or without the help of the user-driven tuning, was able to place relevant items in the higher ranks of the recommended people list.

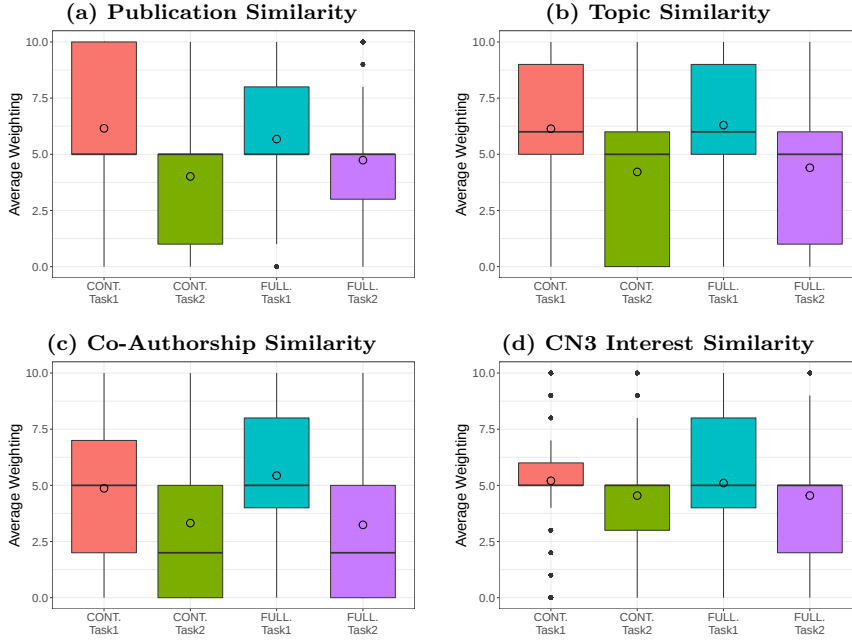


Fig. 4 The average re-weighting value of *CONT* and *FULL* interfaces.

We present the analysis of objective measures in Fig. 5, which shows curves for each measure taken at points $k = 10$ to $k = 100$. A one-way between-subjects ANOVA analysis was conducted to compare the effect of user interfaces on recommendation quality in *BASE*, *CONT*, *EXPL* and *FULL* conditions. In Task 1, we did not find any significant differences between the four conditions. That is, in relevance-oriented Task 1, we did not observe that between the four interfaces, the users selected recommendations from significantly different ranking positions, as measured by TopN, MRR and NDCG metrics. However, we can still see the *FULL* tends to outperform all the other three conditions, as measured by all performance measures. The result hints the controllable and explainable interfaces are useful in increasing the recommendation quality, i.e., the users tend to elevate the more useful recommendations to higher ranks. In Task 1, it is also reasonable to see there is no difference in the alpha-NDCG metric since we did not expect the users to select a set of diverse recommendations.

In Task 2, there was a significant effect of interface on recommendation quality metrics at the $p < 0.05$ level for the four conditions. We found significance on TopN@10 [$F(3, 196) = 5.12, p = 0.001$], MRR@10 [$F(3, 196) = 7.44, p < 0.001$] and NDCG@10 [$F(3, 196) = 7.00, p < 0.001$] metrics. We then conducted a post-test due to the significance. Post-hoc comparisons using a Tukey HSD test indicated that: 1) The TopN mean score for the *FULL* condition ($M = 0.20, SD = 0.14$) was significantly different than the *BASE* ($M = 0.13, SD = 0.15$) and *EXPL* ($M = 0.10, SD = 0.35$) conditions. Moreover, the *CONT* condition ($M = 0.18, SD = 0.13$) did significantly differ from the *EXPL* ($M = 0.10, SD = 0.35$) condition; 2) The MRR mean score for the *FULL* condition ($M = 0.08, SD = 0.06$) was significantly

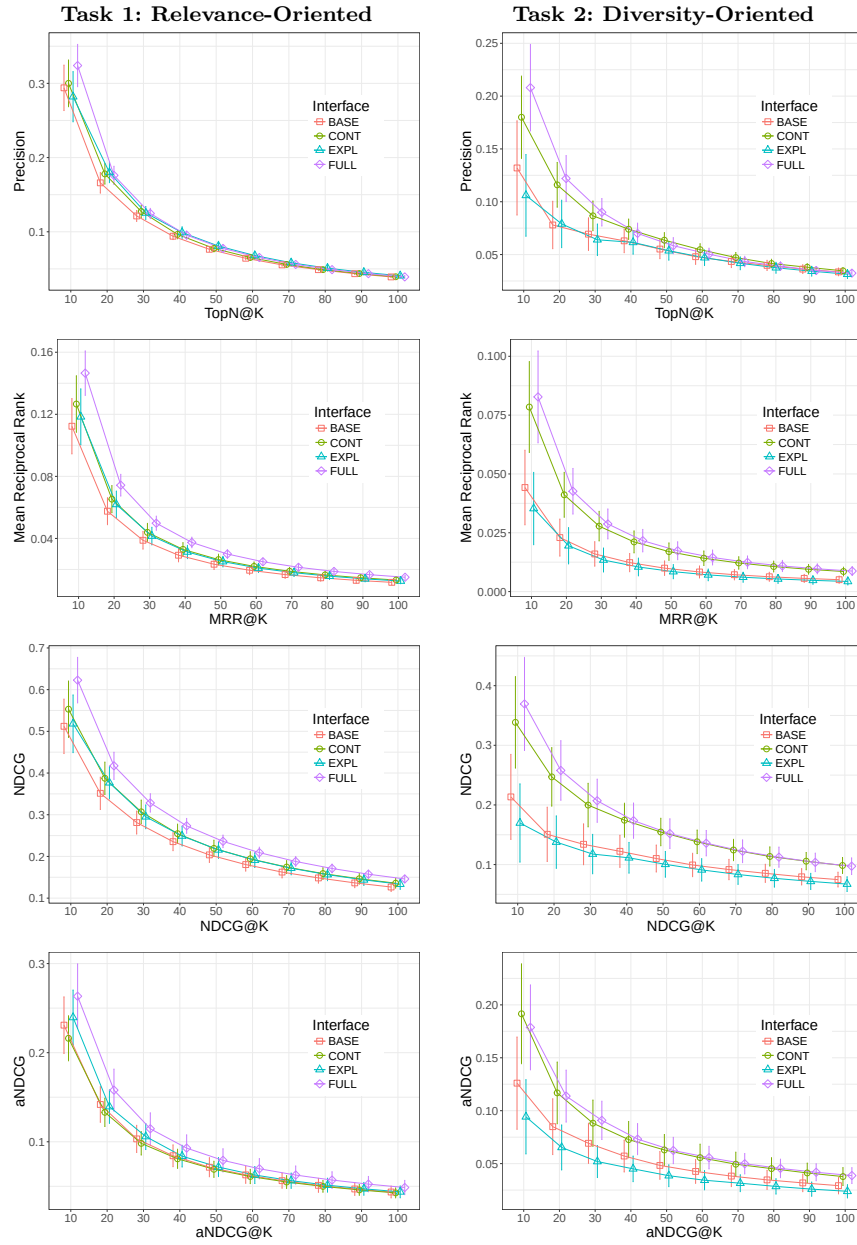


Fig. 5 Objective metrics of recommendation quality: Top-N, MRR, NDCG and alpha-NDCG.

different than the *BASE* ($M = 0.04$, $SD = 0.05$), *CONT* ($M = 0.07$, $SD = 0.06$) and *EXPL* ($M = 0.03$, $SD = 0.5$) conditions. Moreover, the *CONT* condition ($M = 0.07$, $SD = 0.06$) did significantly differ from the *BASE* ($M = 0.04$, $SD = 0.05$) condition; 3) The NDCG mean score for the *FULL* condition ($M = 0.36$, $SD = 0.27$) was significantly different than the *BASE* ($M = 0.21$, $SD = 0.25$) and *EXPL* ($M = 0.16$, $SD = 0.23$) conditions. Moreover, the *CONT* condition ($M = 0.33$, $SD = 0.26$) did significantly differ from the *BASE* ($M = 0.21$, $SD = 0.25$) and *EXPL* ($M = 0.16$, $SD = 0.3$) conditions. We found significance on alpha-NDCG@10 [$F(3, 196) = 4.80$, $p = 0.002$] metric as well. The alpha-NDCG mean score for the *FULL* condition ($M = 0.17$, $SD = 0.14$) was significantly different than the *EXPL* ($M = 0.09$, $SD = 0.12$) conditions. Moreover, the *CONT* condition ($M = 0.19$, $SD = 0.16$) did significantly differ from the *EXPL* ($M = 0.09$, $SD = 0.12$) condition.

To summarize, we found that in diversity-oriented Task 2, providing controllability can increase the overall quality of the recommendation. It means that by using the sliders, users could better tune the ranking to their preferences and bring the most relevant items to the higher positions in the ranked list. However, the findings also indicate that providing explainability may not help to increase the recommendation quality, as measured by our metrics. As we observed, the addition of an explanation function leads the user to go deeper into the ranking list when inspecting and selecting recommendations, which is bad from the perspective of these metrics. We argue, however, that the reduction of quality in the presence of explanations is likely to come from the effect of using traditional ranking-focused recommendation metrics, which were not designed to assess the explanation aspects and might not reflect the true performance of a system nor the ultimate level of user satisfaction. It is known that when decision-support information is scarce, ranking is one of the most important factors to assess item relevance (Keane et al., 2008), causing people to prefer higher-ranked items. However, the presence of decision-support information in the form of explanations could have made the users more confident to explore the lower ranks and judge people positioned there as relevant to meet. In this study, we attempted to minimize the inherent bias of higher ranking by assessing the value of each condition using a combination of objective and subjective metrics.

At the same time, it is interesting to observe that selecting the recommendations beyond the top ranks is not a guarantee that the users will achieve a higher level of diversity within their selection. In the finding of alpha-NDCG, we found a higher alpha-NDCG score in the *CONT* and *FULL* conditions. The *EXPL* demonstrated the lowest alpha-NDCG score, which implied that the users failed to select recommendations that covered sufficiently different research topics.

6.3 User Feedback Analysis

6.3.1 Interface Differences

Table 4 shows the analysis of post-stage survey focused on differences between the interfaces. We built 7 constructs based on 27 questions, including *perceived recommendation quality* (Q), *perceived recommendation diversity* (D), *perceived control* (C), *perceived system effectiveness* (E), *perceived trust* (T), *perceived transparency* (P) and *satisfaction* (S). All these constructs reflected the user experience after

using the four interface conditions. A one-way between subjects ANOVA analysis was conducted to compare the effect of user interfaces on user subjective feedback in the *BASE*, *CONT*, *EXPL*, and *FULL* conditions.

We found significance on all constructs, included *perceived recommendation quality (Q)* [$F(3, 196) = 4.00, p < 0.001$], *perceived recommendation diversity (D)* [$F(3, 196) = 2.96, p = 0.03$], *perceived control (C)* [$F(3, 196) = 12.82, p < 0.001$], *perceived system effectiveness (E)* [$F(3, 196) = 4.62, p < 0.001$], *perceived trust (T)* [$F(3, 196) = 9.77, p < 0.001$], *perceived transparency (P)* [$F(3, 196) = 17.63, p < 0.001$] and *satisfaction (S)* [$F(3, 196) = 6.17, p < 0.001$]. We then conducted a post-test for all constructs that showed significance. The post-hoc comparisons using a Tukey HSD test was reported below:

1. The score of *perceived recommendation quality (Q)* for the *FULL* condition ($M = 4.72$ $SD = 0.68$) was significantly higher than the *BASE* ($M = 4.19$, $SD = 0.91$) condition. However, the *CONT* ($M = 4.52$, $SD = 0.76$) and *EXPL* ($M = 4.50$, $SD = 0.71$) conditions did not significantly differ from the *BASE* condition.
2. The score of *perceived recommendation diversity (D)* for the *FULL* condition ($M = 5.24$ $SD = 0.78$) was significantly higher than the *BASE* ($M = 4.19$, $SD = 0.91$) condition. However, the *CONT* ($M = 5.10$, $SD = 0.97$) and *EXPL* ($M = 5.14$, $SD = 0.78$) conditions did not significantly differ from the *BASE* condition.
3. The score of *perceived control (C)* for the *FULL* ($M = 5.94$ $SD = 0.83$) and *CONT* ($M = 5.66$, $SD = 0.90$) conditions were significantly higher than the *BASE* ($M = 4.84$, $SD = 1.09$) condition. Moreover, it was significantly higher for the *FULL* condition than the *EXPL* ($M = 5.18$ $SD = 1.00$) condition.
4. The score of *perceived system effectiveness (E)* for the *FULL* ($M = 5.04$ $SD = 0.72$) and *CONT* ($M = 4.88$, $SD = 0.69$) conditions were significantly higher than the *BASE* ($M = 4.49$, $SD = 0.87$) condition. Moreover, it was significantly higher for the *FULL* than the *EXPL* ($M = 5.18$ $SD = 1.00$) condition. However, the *EXPL* ($M = 4.83$, $SD = 0.72$) condition did not significantly differ from the *BASE* condition.
5. The score of *perceived trust (T)* for the *FULL* ($M = 5.63$ $SD = 1.13$), *CONT* ($M = 5.25$, $SD = 1.06$) and *EXPL* ($M = 5.32$, $SD = 1.00$) conditions were significantly higher than the *BASE* ($M = 4.84$, $SD = 1.09$) conditions.
6. The score of *perceived transparency (P)* for the *FULL* ($M = 5.74$ $SD = 1.12$), *CONT* ($M = 4.81$, $SD = 1.27$) and *EXPL* ($M = 5.66$, $SD = 1.06$) conditions were significantly higher than the *BASE* ($M = 4.16$, $SD = 1.53$) conditions. Moreover, it was significantly higher for the *FULL* and *EXPL* conditions than the *CONT* condition.
7. The score of *satisfaction (S)* for the *FULL* ($M = 5.60$ $SD = 1.27$) and *CONT* ($M = 5.39$, $SD = 1.06$) conditions was significantly higher than the *BASE* ($M = 4.58$, $SD = 1.4353$) conditions.

To summarize, we found that each of our two novel features, control and explanations, when taken alone, tend to positively affect user experience, as measured by a range of experience-focused constructs, and reached significance for several metrics. As a whole, controllability seems to provide a slightly stronger impact, as it generated higher scores and reached a significant difference with the baseline in more metrics (4 vs. 1). At the same time, the perceived transparency was affected

Construct	Factor	Interfaces			
		BASE M(SD)	CONT M(SD)	EXPL M(SD)	FULL M(SD)
Perceived Recommendation Quality (Q)	Q1	4.86 (1.48)	5.42 (0.99)	5.32 (1.25)	5.72 (1.08)
	Q2	4.96 (1.59)	5.26 (1.27)	5.34 (1.08)	5.76 (1.02)
	Q3	4.48 (1.46)	5.22 (1.25)	5.10 (1.16)	5.48 (1.29)
	Q4	2.46 (1.45)	2.20 (1.41)	2.26 (1.33)	1.92 (1.19)
	Ave	4.19 (0.91)	4.52 (0.76)	4.50 (0.71)	4.72 (0.68)**
Perceived Recommendation Diversity/Variety (D)	D1	5.04 (1.51)	5.48 (1.38)	5.72 (1.12)	6.06 (1.05)
	D2	5.18 (1.20)	5.30 (5.36)	5.32 (1.23)	5.36 (1.54)
	D3	4.86 (1.67)	5.22 (1.44)	5.30 (1.26)	5.46 (1.18)
	D4	3.96 (1.61)	4.42 (1.42)	4.22 (1.63)	4.08 (1.66)
	Ave	4.76 (0.83)	5.10 (0.97)	5.14 (0.82)	5.24 (0.78)**
Perceived Control (C)	C1	4.02 (2.27)	6.02 (1.05)	4.52 (2.09)	6.20 (0.94)
	C2	6.04 (1.17)	6.16 (1.07)	5.82 (1.27)	6.18 (0.91)
	C3	4.70 (1.60)	5.02 (1.54)	5.12 (1.45)	5.60 (1.34)
	C4	4.62 (1.80)	5.46 (1.23)	5.26 (1.39)	5.78 (1.35)
	Ave	4.84 (1.09)	5.66 (0.90)**	5.18 (1.00)	5.94 (0.83)**
Perceived System Effectiveness (E)	E1	4.42 (1.79)	5.26 (1.48)	5.10 (1.31)	5.50 (1.40)
	E2	4.84 (1.53)	5.48 (1.24)	5.60 (1.01)	5.84 (1.21)
	E3	4.92 (1.44)	5.58 (1.24)	5.34 (1.45)	5.82 (1.00)
	E4	3.80 (1.86)	3.20 (1.69)	3.28 (1.69)	3.02 (1.77)
	Ave	4.49 (0.87)	4.88 (0.69)*	4.83 (0.72)	5.04 (0.72)**
Perceived Trust (T)	T1	4.34 (1.67)	5.22 (1.20)	5.18 (1.38)	5.60 (1.24)
	T2	4.40 (1.48)	5.30 (1.24)	5.20 (1.24)	5.52 (1.34)
	T3	4.34 (1.67)	5.02 (1.37)	5.46 (1.14)	5.64 (1.34)
	T4	4.64 (1.39)	5.48 (1.01)	5.44 (1.07)	5.76 (1.28)
	Ave	4.43 (1.40)	5.25 (1.06)	5.32 (1.00)	5.63 (1.13)**
Perceived Transparency (P)	P1	4.08 (1.80)	4.90 (1.51)	5.16 (1.34)	5.58 (1.23)
	P2	3.98 (1.84)	4.44 (1.78)	6.00 (1.17)	5.94 (1.39)
	P3	4.42 (1.72)	5.10 (1.46)	5.82 (1.40)	5.70 (1.40)
	Ave	4.16 (1.53)	4.81 (1.27)*	5.66 (1.06)**	5.74 (1.12)**
Satisfaction (S)	S1	4.46 (1.70)	5.42 (1.27)	5.08 (1.50)	5.74 (1.42)
	S2	4.64 (1.78)	5.38 (1.45)	5.12 (1.43)	5.74 (1.48)
	S3	4.44 (1.57)	5.32 (1.16)	5.20 (1.27)	5.54 (1.32)
	S4	4.80 (1.34)	5.46 (0.99)	5.12 (1.33)	5.40 (1.39)
	Ave	4.58 (1.43)	5.39 (1.06)**	5.13 (1.22)	5.60 (1.27)**

Table 4 User feedback analysis results. We found that the *FULL* condition significantly outperformed the *BASE* condition. Statistical significance level: (*) $p < 0.05$. (**) $p < 0.01$

more significantly by adding explanations. Yet, it is together that these features could provide the strongest impact, producing the highest experience scores and reaching significant difference with the baseline in all experience constructs.

6.3.2 Cohort Differences

It is normal to assume the participants' academic background would affect their perception of using the systems. To examine this aspect of the study, we contrasted user perceptions between two academic groups; i.e. the Ph.D. group and the Masters' group. We found significant differences between the two groups for all constructs, including *perceived recommendation quality (Q)* [$F(1, 192) = 7.56, p < 0.001$], *perceived recommendation diversity (D)* [$F(1, 192) = 11.46, p < 0.001$], *perceived control (C)* [$F(1, 192) = 7.08, p < 0.001$], *perceived system effectiveness (E)* [$F(1, 192) = 4.23, p < 0.001$], *perceived trust (T)* [$F(1, 192) = 20.89,$

$p < 0.001$], *perceived transparency* (P) [$F(1, 192) = 15.13, p < 0.001$] and *satisfaction* (S) [$F(1, 192) = 15.13, p < 0.001$]. We then conducted a post-test for all constructs that shown significance. The post-hoc comparisons using the Tukey HSD test are reported below:

1. The score of *perceived recommendation quality* (Q) for the *Masters'* group ($M = 4.69$ $SD = 0.76$) was significantly higher than the *Ph.D.* ($M = 4.30$, $SD = 0.77$) group.
2. The score of *perceived recommendation diversity* (D) for the *Masters'* group ($M = 5.32$ $SD = 0.81$) was significantly higher than the *Ph.D.* ($M = 4.84$, $SD = 0.86$) group.
3. The score of *perceived control* (C) for the *Masters'* group ($M = 5.61$ $SD = 1.00$) was significantly higher than the *Ph.D.* ($M = 5.23$, $SD = 1.06$) group.
4. The score of *perceived system effectiveness* (E) for the *Masters'* group ($M = 4.97$ $SD = 0.79$) was significantly higher than the *Ph.D.* ($M = 4.67$, $SD = 0.74$) group.
5. The score of *perceived trust* (T) for the *Masters'* group ($M = 5.50$ $SD = 1.17$) was significantly higher than the *Ph.D.* ($M = 4.86$, $SD = 1.21$) group.
6. The score of *perceived transparency* (P) for the *Masters'* group ($M = 5.39$ $SD = 1.32$) was significantly higher than the *Ph.D.* ($M = 4.83$, $SD = 1.44$) group.
7. The score of *satisfaction* (S) for the *Masters'* group ($M = 5.47$ $SD = 1.23$) was significantly higher than the *Ph.D.* ($M = 4.93$, $SD = 1.31$) group.
8. We did not find interaction effect between the interfaces and degree groups for any of the constructs

The result shows a clear pattern: the *Masters'* group had higher user experience scores than the *Ph.D.* group. The findings demonstrate that personal factors (such as level of education, domain experience, and familiarity with technology) could significantly affect the user perception of the system. Given the nature of computer and information science field, the difference in technical knowledge between groups (i.e., interfaces, recommender systems, the Web) is likely to be much smaller than the difference in their domain knowledge (i.e., research, academia, publication, co-authorship, advising). It hints that an interactive and transparent recommender system could be of more value and greater importance to the users with lower domain knowledge for whom making their own decision in a less familiar domain could be a considerable challenge.

7 Structural Equation Modeling

To build a complete understanding of the impact provided by our novel interface features on user experience, we conducted a structural equation model (SEM) analysis advocated in (Knijnenburg et al., 2012). SEM analysis helps to explain the relationship between unobserved constructs (latent variables) using observable variables. For example, we may not be able to find measurements that perfectly represent the user experience of using an intelligent system. However, we can adopt several items (questions) to measure user experience. In this analysis, we build constructs through subjective measurements; we used the post-stage survey comprised of 16 questions that covered 7 different user experience (UX) dimensions.

We also combined 3 constructs of personal characteristics (PCs) and 3 constructs of system-specific characteristics (SCs).

In order to make sure the items (questions) in each construct were meaningful, we started with confirmatory factor analysis (CFA) and examined the construct validity. We tested the construct through two parameters: *convergent validity*, which makes sure the items in the construct are related, and *discriminant validity*, which makes sure the unrelated items are really unrelated. The convergent validity of constructs was maintained by examining the average variance extracted (AVE) of each construct. In our analysis, with one exception, the AVEs of all constructs were higher than the recommended value of 0.50, indicating an adequate convergent validity (Knijnenburg and Willemsen, 2015). We had to remove one item from several constructs due to low variance, but the remaining items shared at least 48% of their variance with their designated construct. In order to ensure discriminant validity, we ascertained that the square root of the AVE for each construct was higher than the highest correlations of the construct with other constructs. We planned three sets of latent constructs: six subjective system aspects (SSA), one user experience (EXP) (satisfaction), three constructs of personal characteristics (PCs), and three constructs of system-specific characteristics (SCs). All statistics summarized below supports good convergent validity (AVE) and internal consistency (Cronbach's α).

- *SSA: perceived recommendation quality (Q)*: 3 items (Q1, Q2, Q3). We removed items Q4 due to low variance (communality: 0.26) with the designated construct. ($AVE = 0.75$, $\sqrt{AVE} = 0.86$, $\alpha = 0.87$, largest correlation = 0.85)
- *SSA: perceived recommendation diversity (D)*: 3 items (D1, D2, D3). We removed item D4 due to low variance (communality: 0.006) with the designated construct. ($AVE = 0.60$, $\sqrt{AVE} = 0.77$, $\alpha = 0.67$, largest correlation = 0.75)
- *SSA: perceived control (C)*: 4 items (C1, C2, C3, C4). While it is a popular construct, in our study the convergent validity was not adequate for this construct ($AVE < 0.5$). We discarded this construct in our analysis. ($AVE = 0.37$, $\sqrt{AVE} = 0.60$, $\alpha = 0.61$, largest correlation = 0.77)
- *SSA: perceived system effectiveness (E)*: 3 items (E1, E2, E3). We removed item E4 due to low variance (communality: 0.31) with the designated construct. ($AVE = 0.75$, $\sqrt{AVE} = 0.86$, $\alpha = 0.86$, largest correlation = 0.75)
- *SSA: perceived trust (T)*: 4 items (T1, T2, T3, T4). ($AVE = 0.76$, $\sqrt{AVE} = 0.86$, $\alpha = 0.91$, largest correlation = 0.85)
- *SSA: perceived transparency (P)*: 3 items (P1, P2, P3). ($AVE = 0.68$, $\sqrt{AVE} = 0.82$, $\alpha = 0.81$, largest correlation = 0.73)
- *EXP: perceived satisfaction (S)*: 4 items (S1, S2, S3, S4). ($AVE = 0.81$, $\sqrt{AVE} = 0.90$, $\alpha = 0.91$, largest correlation = 0.88)

We then built a structural equation model (SEM) for analyzing the UX concepts and directionality of causal effects. We followed the framework recommended in (Knijnenburg and Willemsen, 2015). We iteratively tested and removed the constructs in the SEM model until we found that the model was both stable and fitted. After the analysis, we kept only three constructs: *perceived recommendation quality (Q)*, *perceived transparency (P)*, and *perceived satisfaction (S)* in the model. The model fit the statistics of $\chi^2(200) = 227.29$, $p < 0.001$, $RMSEA = 0.054$, $90\%CI : [0.040, 0.067]$, $CFI = 0.99$, $TLI = 1.00$, which indicates an effective fitting model.

The built model is consistent with findings by (Knijnenburg et al., 2012) that user satisfaction was mediated through the constructs of perceived understandability (in our case, the construct of perceived transparency) and perceived recommendation quality. We believe the selected constructs were essential to our user experiment design; i.e., users were performing information-seeking tasks through the recommender interfaces. At the same time, since the user tasks were not designed to promote recommendation diversity, gain user trust, or let the user make the decision faster (effectiveness), it is not surprising to see that the user experience was not mediated through these constructs. However, it is surprising to see the construct of perceived control had to be removed as well. The results indicated that in our study, user satisfaction was not mediated through the user perception of control; instead, the user satisfaction stems from perceived transparency and perceived recommendation quality.

The model shows that the controllability and explainability manipulations each have an independent positive effect on the *perceived transparency* of the system. The controllable interface is more transparent than the interfaces with no control; however, explainability contributes more to the perceived transparency than controllability. The factor of *perceived transparency* in turn related to users' perception of recommendation quality, which is in turn related to the participants' overall satisfaction with the system. We also found the controllability manipulation has positive effects on the factor of *perceived quality*, but the explainability manipulation has a surprisingly negative effect on the factor of *perceived quality*.

We expanded the core model by adding additional variables of user activity and user characteristics. We found three positive/negative effects of personal characteristic (PCs) on the factor of *perceived transparency*. The finding revealed a higher transparency perception if the participants have lower levels of trust in technology (PC1), have a higher score of scholar expertise (PC2), have a higher acceptance of diversity (of recommendation) (PC3), or have higher privacy concern (SC1). A higher score of acceptance of diversity also has a direct positive effect on the factor of *perceived quality*. Most notably, we found that *system familiarity* has a negative effect on user satisfaction, which indicates that experts may be harder to please. In the variables of user activity, we found that the need to explore more scholars in details (more clicks on scholar name) has a negative effect on the factor of *perceived quality*. At the same time, as the participants work with the system longer (spent more time in solving the tasks), their perception of system transparency increases.

8 Summary and Conclusions

In this paper, we presented a large-scale human subject experiment (N=50) that assessed the impact of controllable and explainable recommender interfaces in a hybrid social recommender system. For the purpose of the study, we augmented a hybrid recommender system, RelevanceTuner+, which offers a user-controllable fusing or recommendation sources through sliders with a total of four explainable source recommendation models (publication, topic, co-authorship, and interest similarities). We then conducted a controlled user study using data from the preceding four years of UMAP conferences. To examine the separate and compound impact of controllability and explainability, we used four conditions: a baseline

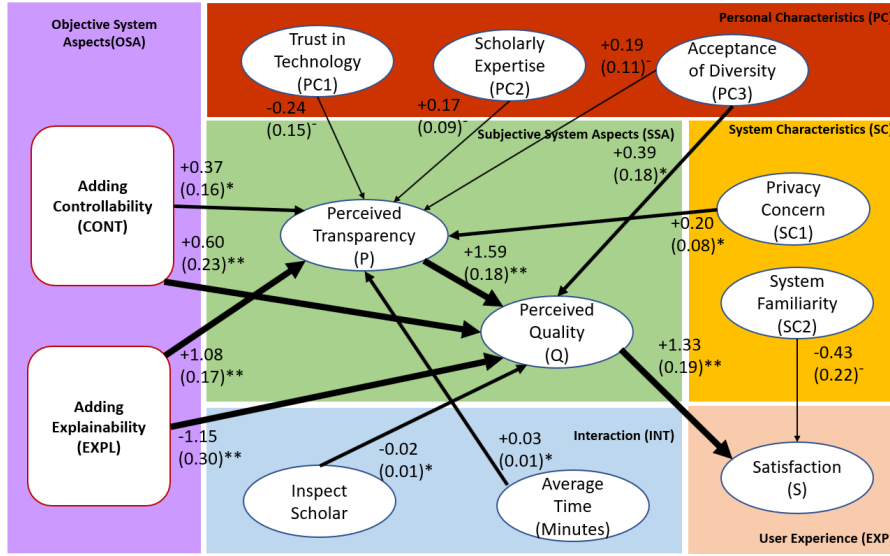


Fig. 6 The structural equation model. The number (thickness) on the arrows represents the coefficients and standard error of the effect. Significance: ** $p < 0.01$, * $p < 0.05$, ⁻ $p < 0.1$.

interface (BASE) with both sliders and access to explanations disabled; a controllable interface (CONT) with slider enabled; an explainable interface (EXPL) with explanation access enabled; and a full interface (FULL) with both slider and explanation access enabled. We followed the within-subject design and assigned the participants two scenario-based tasks, one relevance-oriented and one diversity-oriented. We used a total of 7 constructs with 19 questions as subjective measures. The objective measures were captured by analyzing each user activity log and applying traditional recommendation quality metrics.

The user action analysis demonstrated that the participants extensively adopted the control and explanation functionality that our interfaces provided. We also observed that in their work with sliders, the users *boosted* the relevance weighting in the relevance-oriented task (Task 1) and *decreased* the weighting in the diverse-oriented task (Task 2), as expected by the nature of the tasks.

The analysis of recommendation quality using traditional ranking-focused metrics demonstrated that in the diversity-oriented task, the availability of controllable fusion increased the recommendation quality; that is, the users adopted more recommendations when working with the higher-ranked items. In contrast, providing an explanation function leads the users to click more recommendations beyond top-ranked items, decreasing all ranking-based metrics. Interestingly, the analysis of alpha nDCG demonstrated that picking more items in the lower ranks may not guarantee the high recommendation diversity, which was determined by the dissimilarity between the recommending items.

The results of the subjective feedback analysis demonstrated that the *FULL* condition outperformed all other conditions, which answers RQ1. The analysis shows evidence that the provision of the controllable and explainable user interface has a significant positive effect on user perception. Moreover, the effect of

providing controllable fusion and reasoning transparency is additive. The *FULL* condition outperformed the single enhanced conditions, such as *CONT* and *EXPL*. Another interesting finding indicates the users' academic background affected user perception. We found the participants who are Masters' students were a more "easy-to-please" group. Their subjective feedback score was significantly higher than that of the Ph.D. students.

It is interesting to stress that traditional recommendation quality metrics and user feedback analysis have a seemingly opposing conclusion on the effect of explainability. While controllability had a clear positive impact on user perceptions of the system, it also caused users to select relevant items in lower ranks. It is quite natural to expect that in the presence of explanations, the user has more supporting information for making their choice than with the ranking alone, and would be willing to explore lower-ranked items with increased satisfaction. However, it is a clear sign that traditional ranking-focused metrics must be applied with more caution to interactive recommender systems. In the absence of user-contributed data, the ranking-based analysis might not be sufficient to determine a better design option.

In the section of the structural equation model (SEM) analysis, we used seven subjective constructs, three personal characteristic constructs, and three system-specific characteristic constructs. We first ran a confirmatory factor analysis (CFA) to confirm the convergent validity and convergent validity of constructs. We then included only the valid constructs in the SEM analysis. The results indicate that both controllability (*CONT*) and explainability (*EXPL*) positively contributed to the construct of *perceived transparency (P)*, although the *EXPL* condition had a stronger effect than the *CONT* condition. However, the effect of controllability and explainability on *perceived quality (Q)* was the opposite. The *CONT* condition showed a positive contribution, but the *EXPL* condition negatively contributed to the construct.

We found the best user experience happened when both controllability and explainability were provided (*FULL*). However, we also found the interaction between the two functions. Based on the score of *perceived control (C)*, we found the controllability was additive when the controllable sliders were provided. That is, although explanation function contributed to *perceived control (C)*, but the score can be strengthened by the controllable sliders. In contrast, we found the explanation function contributed to the score of *perceived transparency (P)* more than the controllable slider. That is, although the slider contributed to gain recommender system transparency, the user perception can be improved by providing extra explanations.

This paper confirmed some findings of previous studies while also offering new findings and providing a deeper analysis. Overall, the results support our belief in the effectiveness of the controllable and explainable user interfaces. Moreover, we found the effects of the two explored enhancements are additive, which means the best user experience happens in the *FULL* condition.

9 Limitations and Future Works

One limitation of our work is its scale. While a controlled user study with 50 subjects is considered to be a large scale study, some observed trends might not

be able to reach significance due to their scale. Specifically, a larger number of subjects or communities would be desirable to build a reliable structural equation model of the process. We also noticed that some participants, especially doctoral students, may have highly specific and specialized research interests. In this situation, UMAP conference data used in our study may not offer a sufficient fit to the research interests of all participants. A better match between participants' general interests and the dataset might bring more reliable results. All these limitations are worth further exploration in future study.

In our future work, we also plan to consider the fidelity of explanations and how explanation fidelity influences user satisfaction; i.e., the explanations may not realize the recommendation mechanism with 100% accuracy. The explanations may not be the true reason for the system to make a certain recommendation, or it may only partially realize the reason. It would be exciting to explore how much fidelity can be preserved when conveying the recommendation reasons to users, how much users can actually understand these reasons, and how explanation fidelity will influence user receptiveness and judgment of the recommendations. We also consider that different types of explanations, such as those based on natural language, could have a different impact on user perceptions. This paper explored a similarity-based recommendation model, which is relatively easy to explain. We plan to explore more advanced explainable recommendation models in future user studies. It would be valuable to see how controllability and explainability affects user work with more advanced explainable recommendation approaches. The fidelity problem mentioned above would be even more important in such problem settings. In this way, we may be able to see better integration of machine learning-based and user study-based approaches in the context of controllable/explainable recommendations, which might significantly increase the value of these technologies to end users.

10 Contribution

This paper focuses on the effective design of the user-controllable and explainable interface and examines the value of controllability and explainability in the context of a hybrid social recommender system. The experiment results provide empirical evidence in favor of including these features in interactive recommender systems and uncovers a richer set of dependencies between these features, user interaction, and user experience.

A new contribution of this paper to the area of interactive and transparent recommender systems is a thorough examination of individual and combined effects of controllable and explainable interfaces in a hybrid social recommender system; in particular, its impact on the perceived transparency and satisfaction of the recommendation process. It is also important that our 2-by-2 design allows us to examine the interaction effect between the controllable and explainable interfaces. The design can be easily extended to apply to many different contexts.

Our contribution covers several aspects of transparency of recommendation. First, we propose and evaluate a novel user-controllable interface for hybrid recommender systems and design a range of explanatory visualizations to enhance the controllability and explainability of a social recommender system. Second, we discuss the trade-off between the transparent fusion and reasoning processes,

which implies the interaction effects between controllability and explainability in a social recommender system. Third, we present implications for the design of the user interface to enhance user controllability and explainability based on these results. Our work has a potential to extend to other recommender systems beyond this context, thereby making significant contributions to the research topics on the intelligent user interface (IUI); fair, accountable, and transparent (FAT) recommender systems (RS); and the domain of explainable AI (XAI).

References

- Abdollahpouri H, Adomavicius G, Burke R, Guy I, Jannach D, Kamishima T, Krasnodebski J, Pizzato L (2019) Beyond personalization: Research directions in multistakeholder recommendation. arXiv preprint arXiv:190501986
- Ahn Jw, Brusilovsky P, Grady J, He D, Florian R (2010) Semantic annotation based exploratory search for information analysts. *Information processing & management* 46(4):383–402
- Amershi S, Weld D, Vorvoreanu M, Fournery A, Nushi B, Collisson P, Suh J, Iqbal S, Bennett PN, Inkpen K, et al. (2019) Guidelines for human-ai interaction. In: *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, ACM, p 3
- Baeza-Yates R, Ribeiro-Neto B, et al. (1999) *Modern information retrieval*, vol 463. ACM press New York
- Bailey SM, Wei JA, Wang C, Parra D, Brusilovsky P (2018) Cnvis: A web-based visual analytics tool for exploring conference navigator data. *Electronic Imaging* 2018(1):1–11
- Bogers T (2018) Tag-based recommendation. In: *Social Information Access*, Springer, pp 441–479
- Bostandjiev S, O'Donovan J, Höllerer T (2012) Tasteweights: a visual interactive hybrid recommender system. In: *Proceedings of the sixth ACM conference on Recommender systems*, ACM, pp 35–42
- Brusilovsky P, Oh JS, López C, Parra D, Jeng W (2016) Linking information and people in a social system for academic conferences. *New Review of Hypermedia and Multimedia* pp 1–31
- Burke R (2007) Hybrid web recommender systems. In: *The Adaptive Web: Methods and Strategies of Web Personalization*, Springer-Verlag
- Chen J, Geyer W, Dugan C, Muller M, Guy I (2009) Make new friends, but keep the old: recommending people on social networking sites. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ACM, pp 201–210
- Chen L, Wang F (2017) Explaining recommendations based on feature sentiments in product reviews. In: *Proceedings of the 22Nd International Conference on Intelligent User Interfaces*, ACM, New York, NY, USA, IUI '17, pp 17–28, DOI 10.1145/3025171.3025173
- Clarke CL, Kolla M, Cormack GV, Vechtomova O, Ashkan A, Büttcher S, MacKinnon I (2008) Novelty and diversity in information retrieval evaluation. In: *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, ACM, pp 659–666
- Cramer H, Evers V, Ramlal S, Van Someren M, Rutledge L, Stash N, Aroyo L, Wielinga B (2008) The effects of transparency on trust in and acceptance of a content-based art recommender. *User Modeling and User-Adapted Interaction* 18(5):455
- Eiband M, Schneider H, Bilandzic M, Fazekas-Con J, Haug M, Hussmann H (2018) Bringing transparency design into practice. In: *23rd International Conference on Intelligent User Interfaces*, ACM, pp 211–223
- Ekstrand MD, Kluver D, Harper FM, Konstan JA (2015) Letting users choose recommender algorithms: An experimental study. In: *Proceedings of the 9th ACM Conference on Recommender Systems*, ACM, pp 11–18

- Gedikli F, Jannach D, Ge M (2014) How should i explain? a comparison of different explanation types for recommender systems. *International Journal of Human-Computer Studies* 72(4):367–382
- Gretarsson B, O'Donovan J, Bostandjiev S, Hall C, Höllerer T (2010) Smallworlds: visualizing social recommendations. In: *Computer Graphics Forum*, Wiley Online Library, pp 833–842
- Guy I (2015) Social recommender systems. In: *Recommender Systems Handbook*, Springer, pp 511–543
- Guy I (2018) People recommendation on social media. In: *Social Information Access*, Springer, pp 570–623
- Guy I, Ronen I, Wilcox E (2009) Do you know?: recommending people to invite into your social network. In: *Proceedings of the 14th international conference on Intelligent user interfaces*, ACM, pp 77–86
- Han S, He D, Jiang J, Yue Z (2013) Supporting exploratory people search: a study of factor transparency and user control. In: *Proceedings of the 22nd ACM international conference on Information & Knowledge Management*, ACM, pp 449–458
- Harper FM, Xu F, Kaur H, Condiff K, Chang S, Terveen L (2015) Putting users in control of their recommendations. In: *Proceedings of the 9th ACM Conference on Recommender Systems*, ACM, pp 3–10
- He C, Parra D, Verbert K (2016) Interactive recommender systems: a survey of the state of the art and future research challenges and opportunities. *Expert Systems with Applications* 56:9–27
- Herlocker JL, Konstan JA, Riedl J (2000) Explaining collaborative filtering recommendations. In: *Proceedings of the 2000 ACM conference on Computer supported cooperative work*, ACM, pp 241–250
- Jannach D, Naveed S, Jugovac M (2016) User control in recommender systems: Overview and interaction challenges. In: *International Conference on Electronic Commerce and Web Technologies*, Springer, pp 21–33
- Keane M, O'Brien M, Smyth B (2008) Are people biased in their use of search engines? *Commun ACM* 51(2):49–52
- Kluser D, Ekstrand MD, Konstan JA (2018) Rating-based collaborative filtering: algorithms and evaluation. In: *Social Information Access*, Springer, pp 344–390
- Knijnenburg BP, Willemsen MC (2015) Evaluating recommender systems with user experiments. In: *Recommender Systems Handbook*, Springer, pp 309–352
- Knijnenburg BP, Bostandjiev S, O'Donovan J, Kobsa A (2012) Inspectability and control in social recommenders. In: *Proceedings of the sixth ACM conference on Recommender systems*, ACM, pp 43–50
- Kobsa A, Koenemann J, Pohl W (2001) Personalised hypermedia presentation techniques for improving online customer relationships. *The knowledge engineering review* 16(2):111–155
- Kouki P, Schaffer J, Pujara J, O'Donovan J, Getoor L (2017) User preferences for hybrid explanations. In: *Proceedings of the Eleventh ACM Conference on Recommender Systems*, ACM, pp 84–88
- Lee D, Brusilovsky P (2018) Recommendations based on social links. In: *Social Information Access*, Springer, pp 391–440
- Manning C, Raghavan P, Schütze H (2010) Introduction to information retrieval. *Natural Language Engineering* 16(1):100–103

- Musto C, Narducci F, Lops P, de Gemmis M, Semeraro G (2019a) Linked open data-based explanations for transparent recommender systems. *International Journal of Human-Computer Studies* 121:93–107
- Musto C, Rossiello G, de Gemmis M, Lops P, Semeraro G (2019b) Combining text summarization and aspect-based sentiment analysis of users' reviews to justify recommendations. In: *Proceedings of the 13th ACM Conference on Recommender Systems*, pp 383–387
- O'Donovan J, Smyth B, Gretarsson B, Bostandjiev S, Höllerer T (2008) Peer-chooser: visual interactive recommendation. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ACM, pp 1085–1088
- O'Donovan J, Gretarsson B, Bostandjiev S, Hollerer T, Smyth B (2009) A visual interface for social information filtering. In: *Computational Science and Engineering, 2009. CSE'09. International Conference on*, IEEE, vol 4, pp 74–81
- O'Mahony MP, Smyth B (2018) From opinions to recommendations. In: *Social Information Access*, Springer, pp 480–509
- Papadimitriou A, Symeonidis P, Manolopoulos Y (2012) A generalized taxonomy of explanations styles for traditional and social recommender systems. *Data Mining and Knowledge Discovery* 24(3):555–583
- Parra D, Brusilovsky P (2015) User-controllable personalization: A case study with setfusion. *International Journal of Human-Computer Studies* 78:43–67
- Pu P, Chen L (2007) Trust-inspiring explanation interfaces for recommender systems. *Knowledge-Based Systems* 20(6):542–556
- Pu P, Chen L, Hu R (2011) A user-centric evaluation framework for recommender systems. In: *Proceedings of the fifth ACM conference on Recommender systems*, ACM, pp 157–164
- Schafer JB, Konstan JA, Riedl J (2002) Meta-recommendation systems: user-controlled integration of diverse recommendations. In: *Proceedings of the eleventh international conference on Information and knowledge management*, ACM, pp 43–51
- Schaffer J, O'Donovan J, Michaelis J, Raglin A, Höllerer T (2019) I can do better than your ai: expertise and explanations. In: *IUI*, pp 240–251
- di Sciascio C, Sabol V, Veas EE (2016) Rank as you go: User-driven exploration of search results. In: *Proceedings of the 21st International Conference on Intelligent User Interfaces*, ACM, pp 118–129
- di Sciascio C, Brusilovsky P, Veas E (2018) A study on user-controllable social exploratory search. In: *International Conference on Intelligent User Interfaces*
- Sharma A, Cosley D (2013) Do social explanations work?: studying and modeling the effects of social explanations in recommender systems. In: *Proceedings of the 22nd international conference on World Wide Web*, ACM, pp 1133–1144
- Sie RL, Drachslar H, Bitter-Rijkema M, Sloep P (2012) To whom and why should i connect? co-author recommendation based on powerful and similar peers. *Int J Technology Enhanced Learning* 4(1/2):121–137
- Swearingen K, Sinha R (2001) Beyond algorithms: An hci perspective on recommender systems. In: *ACM SIGIR 2001 Workshop on Recommender Systems*
- Tintarev N, Masthoff J (2011) Designing and evaluating explanations for recommender systems. In: *Recommender systems handbook*, Springer, pp 479–510
- Tintarev N, Masthoff J (2012) Evaluating the effectiveness of explanations for recommender systems. *User Modeling and User-Adapted Interaction* 22(4-5)

- Tintarev N, Masthoff J (2015) Explaining recommendations: Design and evaluation. In: *Recommender Systems Handbook*, Springer, pp 353–382
- Tsai CH, Brusilovsky P (2017) Providing control and transparency in a social recommender system for academic conferences. In: *Proceedings of the 25th Conference on User Modeling, Adaptation and Personalization*, ACM, pp 313–317
- Tsai CH, Brusilovsky P (2018) Beyond the ranked list: User-driven exploration and diversification of social recommendation. In: *23rd International Conference on Intelligent User Interfaces*, IUI '18, pp 239–250
- Tsai CH, Brusilovsky P (2019a) Designing explanation interfaces for transparency and beyond. In: *Workshop on Intelligent User Interfaces for Algorithmic Transparency in Emerging Technologies*
- Tsai CH, Brusilovsky P (2019b) Evaluating visual explanations for similarity-based recommendations: User perception and performance. In: *Proceedings of the 27th ACM Conference on User Modeling, Adaptation and Personalization*, ACM, pp 22–30
- Tsai CH, Brusilovsky P (2019c) Explaining recommendations in an interactive hybrid social recommender. In: *Proceedings of the 24th International Conference on Intelligent User Interfaces*, ACM, pp 391–396
- Verbert K, Parra D, Brusilovsky P, Duval E (2013) Visualizing recommendations to support exploration, transparency and controllability. In: *Proceedings of the 2013 international conference on Intelligent user interfaces*, ACM, pp 351–362
- Vig J, Sen S, Riedl J (2009) Tagsplanations: explaining recommendations using tags. In: *Proceedings of the 14th international conference on Intelligent user interfaces*, ACM, pp 47–56
- Wang W, Benbasat I (2007) Recommendation agents for electronic commerce: Effects of explanation facilities on trusting beliefs. *Journal of Management Information Systems* 23(4):217–246
- Wu Y, Ester M (2015) Flame: A probabilistic model combining aspect based opinion mining and collaborative filtering. In: *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*, ACM, New York, NY, USA, WSDM '15, pp 199–208, DOI 10.1145/2684822.2685291

Chun-Hua Tsai is an assistant research professor of Information Sciences and Technology at Pennsylvania State University. His specialization is in human-computer interaction (HCI) and data science. Chun-Hua Tsai has extensive research experience in recommender systems, user intelligent interfaces, social computing, accessibility and health informatics. He published numerous conference and journal papers in the HCI and AI community.

Peter Brusilovsky is a Professor of Information Science and Intelligent Systems at the University of Pittsburgh, where he directs Personalized Adaptive Web Systems (PAWS) lab. Peter Brusilovsky has been working in the field of adaptive educational systems, user modeling, and intelligent user interfaces for more than 30 years. He published numerous papers and edited several books on adaptive hypermedia, adaptive educational systems, user modeling, and the adaptive Web.