Noname manuscript No.
(will be inserted by the editor)

The Role of Transparency in Multi-Stakeholder Educational Recommendations

Yong Zheng[1] · **Juan Ruiz Toribio**[1,2]

Received: date / Accepted: date

Abstract Recommender systems have been successfully applied to alleviate the problem of information overload and assist users' decision makings. Multi-stakeholder recommender systems produce the item recommendations to the end user by considering the perspective of multiple stakeholders. Existing research on multi-stakeholder recommendations rely on the offline evaluations only, the online studies is still under investigation. This paper made the first attempt to examine the multi-stakeholder recommendations through online studies. On one hand, we use online user studies to compare different recommendation models. On the other hand, we develop novel user interfaces to enhance the transparency, and exploit the role of transparency in multi-stakeholder recommender systems. We collect our own dataset in educational learning and use it as case study to perform the online studies in multi-stakeholder recommendations. We observe that the explanation of the key parameters in the recommendation models can enhance the transparency, which further affects the decision-making of different stakeholders.

Keywords recommender system · multi-stakeholder · utility · transparency · user studies

Yong Zheng
E-mail: yzheng66@iit.edu

Juan Ruiz Toribio
E-mail: jruiztoribio@hawk.iit.edu

1. Illinois Institute of Technology, USA
2. Universidad Politécnica de Madrid, Spain

1 Introduction

Recommender systems (RS) produce item recommendations tailored to user preferences. The end user is the only stakeholder in the traditional RS. Recently, researchers claim that the perspective of other stakeholders are also important, and it is necessary to balance the needs of multiple stakeholders in the recommendation process [9]. For example, there are three stakeholders as shown in Figure 1 – the *advertising agency* would like to present the car advertisement to any *viewers* who may click it. The receiver of the advertisement may just want to view the ones that they are interested in. However, the advertisements should be delivered to potential customers from the perspective of *car producers* or *product sellers*. Teenagers may like cars, but they may not have the capability to make purchases, which decreases the utility of advertisements from the perspective of the sellers. Maximizing the preference of the end users may hurt the utility of the item from the perspective of other stakeholders. Multi-stakeholder recommender systems (MSRS), therefore, suggest producing recommendations by considering multiple stakeholders, and balance their needs in the environment.

Fig. 1: Example in Car Advertising

Researchers have made their attempts to develop different solutions for the multi-stakeholder recommendations in several domains. Burke et al. [9] produced the movie recommendations by considering the movie watchers, suppliers, and the platform. Nguyen et al. [38] utilized learning-to-rank to optimize the objectives from the room suppliers, the consumers and the hotel agency Expedia, Inc. Louca et al. [33] proposed a joint optimization method to consider the recommendation performance, as well as the profits for the platform. Our previous work used the Pareto multi-objective learning to deliver multi-stakeholder recommendations in the online dating [58] and educational domains [64].

Going beyond the algorithmic aspect of the recommendations, the importance of the user-centric evaluations [42] and the explanations of recommendations [36,50] were raised in the research community of RS. However, the existing work on MSRS only deliver offline experiments and evaluations. In this paper, we make the first attempt to perform online studies. We have two major reasons or motivations to investigate the online experiments for MSRS. First, we believe the online evaluation is necessary to balance the needs of multiple stakeholders in MSRS. Without online studies, we do not know a 50%-50% balance is enough for two stakeholders, or other degree of the balance is preferred. In addition, most RS act like black boxes with low transparency in which people do not know the inner operations of the algorithms due to that this information is proprietary and/or sufficiently complex to not be understood [45]. The recommendations may be delivered to the users without explaining how the system works and how the decisions were made, not to mention exposing parameters in the algorithmic process and letting the users tune up the recommendations. The situation could be even worse for multi-stakeholder recommendations. It is because maximizing the preferences of a stakeholder may hurt other stakeholders, especially when there are conflicting interests among these stakeholders. As a result, enhancing transparency is one method to disclose more information, deliver more explanations, and improve the trust of the users in the RS [47].

In our previous work, we have made the attempt to build MSRS for the course project recommendations in the domain of education [64,59]. However, these works only evaluate the models by using offline experiments. It may be difficult to examine whether the model can achieve an actual balance among different stakeholders by using the offline experiments only. In this paper, we focus on the online user studies. There are two issues or challenges in the online evaluations for MSRS – Traditional RS collect user studies from the perspective of the end users only, but we have multiple stakeholders in MSRS. It is necessary to conduct the user studies for each stakeholder. Moreover, to achieve the balance of the needs among different stakeholders, we need to let different stakeholders be aware of this objective in the online studies. In this case, the explanation and transparency in RS are crucial in the experimental design. In this paper, we select course project recommendation in the area of educations as a case study, build multi-stakeholder recommendation models, and perform both offline and online experiments. Particularly, we design novel user interfaces to enhance the transparency in MSRS.

The major contributions in this paper can be summarized as follows.

- It is the first attempt to analyze the role of transparency in multi-stakeholder recommendations.
- It is the first attempt to evaluate multi-stakeholder recommendations by using both offline experiments and user studies.
- Traditional methods to enhance the transparency rely on the explanation of the recommendations, such as why we recommend a specific item to the users. In our paper, we propose a new method in which we disclose and explain the role of key parameters in the recommendation models to the

users and observe how different stakeholders react on the changes of the parameters.
- In addition, we develop an interactive user interfaces which let the users tune up the key parameters in the recommendation process.

The remainder of this paper is organized as follows. Section 2 positions the related work. Section 3 presents the background of our educational case study and introduces how we collected the rating data for offline experiments. Section 4 discusses our previous work on utility-based multi-stakeholder recommendation models. Section 5 presents the offline and online experimental design and protocols. We discuss the experimental results, findings, limitations, and lessons learned in Section 6, followed by the conclusions and future work in Section 7.

2 Related Work

Since our work is relevant to multiple aspects of the RS, we introduce and discuss the following related work.

2.1 Utility-based Recommender Systems

In the past decades, several recommendation models have been developed to produce more accurate recommendations. According to the classification of RS by Burke [7], there are five categories – collaborative models, content-based recommenders, methods which utilize demographic information, knowledge-based algorithms, and utility-based models. The utility-based recommenders make suggestions based on a computation of the utility of each item for the user. Utility can be used to indicate how valuable an item is from the perspective of a user. The utility function may vary from data to data, and there is no unified function to be generalized to different domains or applications. Guttman used different transformation functions (e.g., linear, square, or universal functions) for different types of the attributes (e.g., continuous or discrete) in the context of online shopping [21]. Li et al. [30] defined the utility of recommending a potential link in the social networks by a linear aggregation of its value, cost, and the linkage likelihood. Moreover, Zihayat et al. proposed to use the aggregation of article-driven (e.g., popularity, topic distributions) and user-driven measures (e.g., clickstream, dwell time) as the utility function for news recommendations [65]. Deng [13] considered all the aggregation functions (e.g., linear or radial basis functions) in the multi-criteria RS [31,2,23] as the utility functions. Recently, we took advantage of the multi-criteria ratings and utilized the similarity or distance between the expectation and rating vectors as the utility functions [61]. These expectation and rating vectors are mapped to the same dimensions, i.e., different aspects of the items, such as location, room size and room cleanliness in hotel bookings. Once the utility function is built, the score produced by the function for a user and an item can be used to rank the candidate items, in order to deliver the top-N recommendations to the user.

2.2 Multi-Stakeholder Recommender Systems

In the traditional RS, the receiver of the recommendations is the only stakeholder since the system usually produces the list of recommended items by considering the user preferences only. However, other stakeholders may also play an important role in the process. Actually, the idea of "multiple stakeholders" is not novel, and we can find its trace in the past research on reciprocal recommendations and the marketplace. More specifically, the reciprocal recommendations, dating [6,41,58] and recruitment [55,4], usually consider a bilateral relationship. Take job seeking for example, a good recommender cannot simply suggest all the job positions that a user prefers. The recommender should also consider whether the employer would like to hire the user, in order to improve the probability of a match and avoid frustrations in job seeking. Furthermore, marketplace [5,25,11] is also a good example, especially when it comes to online auctions, such as eBay. There are at least two stakeholders – buyers and sellers. The buyers would like to purchase the item at an affordable price, while the sellers are willing to obtain more profits by selling items at a higher price. It can be extended to the environment with more stakeholders if we consider the marketplace platform [33] and the stakeholders in the supply chains [53]. In addition, one of the research problems in the area of RS is how to achieve the balance between the recommendation accuracy and the platform benefits (e.g., profits) [3,33]. These works can also be considered as multi-stakeholder recommendations since they consider the perspectives from two stakeholders, i.e., the platform and the end users.

MSRS were formally proposed in 2016 [9]. MSRS tries to produce the item recommendations by considering multiple stakeholders, in order to balance the needs of multiple stakeholders. Researchers have discussed the usefulness and necessity of applying MSRS to different domains, such as advertising and movies [9], marketplace [38], educations [16,8,64,59], and so forth. The major characteristics of MSRS can be summarized as follows.

- There are at least two stakeholders in the system, and these stakeholders must have underlying interactions or connections. Take the dating application for example, the relationship between the target user and the partners to be recommended is reciprocal or bilateral. The target user can select any partners, while he or she can be selected by others as well.
- There may be conflicts of interests among different stakeholders. As a result, maximizing the preference of one stakeholder may hurt the utility of the item from the perspective of other stakeholders.

Researchers have developed different solutions to balance the needs of several stakeholders. The earlier work incorporated the notion of multiple stakeholders and objectives but failed to build the models by using optimizations. For example, Chen et al. used the product of purchase probability and item profitability as the score to rank items [11]. The later research work became consistent, most of which utilized the multi-objective learning to solve the problems in MSRS. These works can be summarized as a three-step unified workflow – First of all, we need to figure out the stakeholders in the applications or domains. Afterwards, we can define the utility functions and objectives for each stakeholder. The utility function indicates how valuable a single item is or a list of items are from the perspective of each stakeholder. For example, the rating prediction function by a user on an item can be used to tell the utility of an item from the perspective of the user. Alternatively, the function of calculating precision by given a recommendation list can be used to indicate the utility of the recommendation list from the perspective of the user. Finally, the multi-stakeholder recommendation problem can be transformed to a multi-objective optimization problem in which we can produce the top-N recommendations by achieving the balance among the objectives in view of the different stakeholders.

According to the existing research on multi-stakeholder recommendations, there are generally three optimization methods that can be used for multi-objective learning:

- *Convert multiple objectives into a single one.* For example, Nguyen et al. [38] and Louca et al. [33] used a linear aggregation function to fuse the multiple objectives from different stakeholders and built learning-to-rank models by learning the weights associated with multi-objectives in the aggregation function.
- *Optimize a single objective by setting others as constraints.* For example, Sürer et al. [48] developed a constraint-based integer programming optimization model, in which they can produce the top-N item recommendations for each user and satisfy certain provider constraints.
- *Use the Pareto multi-objective optimization.* These optimizers can learn the solutions and finally produce a Pareto optimal set that is a set of optimal solutions in which no single objective can be further improved without hurting others. The remaining challenge is how to select a single solution from this Pareto optimal set. Lin et al. used the Pareto optimizer for multi-objective recommendations, and they adopted the "least misery" strategy which minimizes the highest loss among multiple objectives [32]. Our previous work calculated an overall loss by aggregating the losses in different objectives, and selected the solution with minimal overall loss in multi-stakeholder recommendations [64, 60].

In terms of the evaluations in MSRS, the existing work focused on offline experiments only. More specifically, researchers can build individual recommendation model which considers the utility of the items from the perspective of each stakeholder, in addition to the multi-stakeholder recommendation models. Afterwards, the objectives from different stakeholders based these recommendation models can be compared, while we can observe the loss and gain, as well as the degree of balance among these objectives. However, the degree of balance is difficult to be defined in advance or achieved by the optimal solutions. Without user studies or online experiments, it is difficult to find the best cut-off value towards the balance.

2.3 Explanation and Transparency in Recommender Systems

The importance of FAT (i.e., Fairness [40,10,12], Accountability [40], and Transparency [47]) has been realized in the area of recommender systems. It was also mentioned as one of the research directions in MSRS [1]. In this paper, we only focus on the aspect of transparency.

The definition and the usefulness of transparency may vary from domains to domains. Lyons defines transparency as a method for establishing shared awareness and shared intent between humans and machines [34], and exploits its impact on human-robot interactions. Deuze believes that transparency allows audience to see more of the news productions and the journalists behind the new stories [14]. The development of computational or algorithmic journalism which employs algorithms to generate news stories with little human interventions may build a barrier for transparency in the news media [15,27]. From the perspective of the applications which utilize data mining, machine learning or artificial intelligence techniques [46,49,43], the transparency mechanisms provide the opportunities for users to gain familiarity with aspects of a system that may be hidden [19]. The importance of transparency has been recognized in expert systems [46], visualizations [44], machine learning [49] and RS [47,39,37]. With greater transparency, we can enable trust of automated systems [35,39], allow people to question, critique and give suggestions in a system [56], provide explanations and help understand the process of algorithmic decision-making systems [29,50], and so forth.

In the area of RS, it is well-known that not only the offline evaluation metrics (e.g., accuracy) but also the subjective user experiences are important. The user-centric evaluation [42], therefore, was promoted for RS. Researchers believe that the explanation of recommendations can be considered as one approach to improve the recommendations and user experiences. Tintarev et al. [50] derived seven aims of the explanation facility which can help a recommender significantly improve user experiences – *transparency* (explains why recommendations were generated), *scrutability* (the ability for the user to critique the system), *trust* (increase users' confidence in the system), *effectiveness* (help users make good decisions), *persuasiveness* (convince users to try or buy items recommended to them), *efficiency* (help users make decisions faster) and *satisfaction* (increase the ease of use and enjoyment).

Sinha et al. point out that the major role of transparency in RS is enabling user understanding of why a particular recommendation was made to the users [47]. The transparency further allows the user to initiate a predictable and efficient interaction with the system [24,18], increase the perceived value and overall acceptance of RS [28], improve trusts on the RS [22,40,39], and so forth.

One of the challenges in RS is how to enhance and evaluate transparency. There are two major solutions in the existing work – providing useful explanations on the user interface or designing interactive user interface in the systems. For example, Musto et al. [37] integrated the framework of explanations in a conversational recommender system in which the user interface can present natural language explanations to tell why the users might like the recommended items. An example of the explanations in the movie domain is, “I suggest you The movie *Untouchables* because you like movies where: the director is Brian De Palma as in Causality of War, the genre is Thriller as in The Departed.” They evaluated the transparency by deploying questionnaires to the end users. The examples of the questions are, “I understood why this movie was recommended to me”, “The explanation increased my trust in RS”, “The explanation made the recommendation more convincing”, etc. They concluded that an effective explanation may help the user to evaluate to what extent she would like the recommendation.

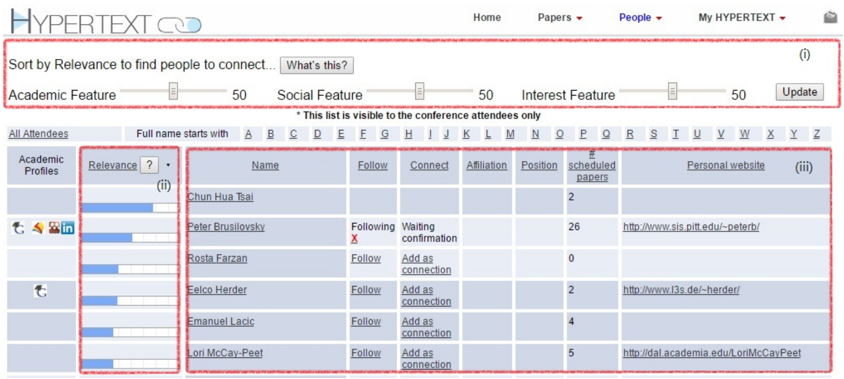


Fig. 2: Interactive User Interface on RelExplorer [51]

By contrast, Tsai et al. [51] developed RelExplorer which recommends and ranks the co-attendees at an academic conference. Particularly, the interactive user interface on RelExplorer (as shown by Figure 2) allows the users to adjust the weights of three recommendation factors – academic feature (e.g., similarity in publications), social feature (e.g., co-author relationships) and interest feature (e.g., co-bookmarked papers) by dragging the slider bars associated with each factor. Every time when a user changes the weights, the list of recommended co-attendees will be re-ranked and displayed on the interface. They additionally present different visualizations (e.g., a word cloud based on publication similarities) as the explanation components to enhance transparency. They used the post-study questionnaire to evaluate the effectiveness of the interactive design and explanation components, e.g., let a user rates whether the user controls can increase the variety of exploration, and whether the visualizations can help them understand how the recommendations were produced.

Using questionnaires to collect the subjective options is the most popular strategy to examine the transparency in RS. However, there are several disadvantages of using the subjective questionnaires, such as limited response rates, validity, reliability, sample biases, and so forth [17, 54, 20]. Inspired by the example given by Figure 2, we believe the interactive user interface is a better way to collect respondents' behaviors. By this way, we can observe the subjects' reactions which can be used to infer the impact of the transparency.

3 Educational Setting and Data

In this section, we introduce the background of our educational case study, as well as the dataset we have for the purpose of offline experiments.

We collect our own data in the process of academic teaching and learning [57]. We collected the data from a Web-based learning portal that improves the process of teaching and learning for instructors and students. One of the components in our portal is the project recommendations. The graduate students who enrolled in the data analytics course are required to complete a final project at the end of the semester. They need to find a dataset from Kaggle.com, define research problems (e.g., hypothesis testing, regressions, classifications, etc.), and utilize their skills to solve the proposed problems.

We design a questionnaire in which we provide a list of potential Kaggle datasets and collect student preferences on them. In the questionnaire, each student should select at least three liked and disliked datasets or projects from a list of 70 pre-defined Kaggle dataset and give an overall rating to these selected ones. In addition, they were asked to rate each selected project on three criteria: how interesting the application area is (*App*), how convenient or easy the data processing will be (*Data*), how easy the whole project is (*Ease*) by using this dataset. The rating scale for all ratings is 1 to 5. The dimension "Data" and "Ease" indicate the degree of ease of the project from different perspectives. Table 1 presents an example of our data.

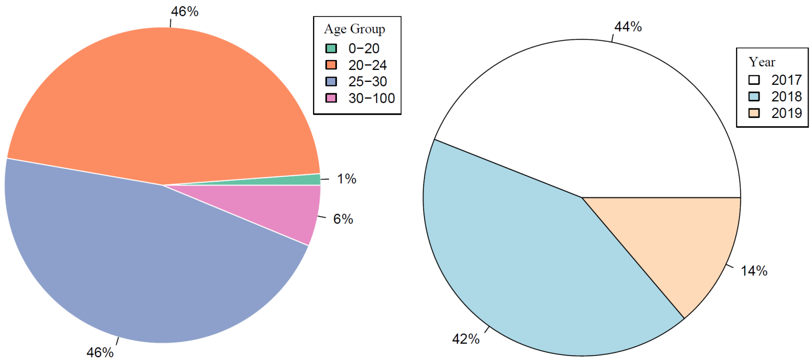
Table 1: Example of The Educational Data

Student	**Item**	**Overall Rating**	**App**	**Data**	**Ease**
10	41	4	4	4	4
10	60	2	2	2	2
12	21	4	4	5	4
...	...	...	...	...	...

We assign this questionnaire to the students in the data analytics class every semester. At the moment when the students were asked to fill the questionnaire, it is usually the 9^{th} or 10^{th} of semester[1], and they have already learned data preprocessing, descriptive statistics, hypothesis testing, linear regression and logistic regressions. There is a total of 3,306 rating entries given by 269 students on 70 Kaggle datasets. The density of the rating data is 17.56%. Each rating entry is composed of an overall rating and multi-criteria ratings by a student on a selected item. More information about the data characteristics can be

[1] There are usually 16 weeks in a semester.

shown by Figure 3 and Figure 4. Among these 269 students, 40.6% of them are males, and 59.4% are females. Moreover, 46% and 46% of the students fall in the age group of [20, 24] and [25, 30] respectively. We have collected the data for five semesters from 2017 to 2019, while the distribution of the ratings over years can be observed from Figure 3 b). In addition, the comparison of the ratings and the correlation among ratings can be shown in Figure 4. We use "Rating" to indicate the overall rating on the items, while "App", "Data" and "Ease" refer to the multi-criteria ratings. Recall that the ratings on "Data" and "Ease" can tell the degree of ease from the perspective of the students. Some students prefer to select easier projects, while others may choose more challenging ones. That is the reason why the correlation between the overall rating and ratings on Data or Ease is lower than the correlation between the overall ratings and ratings on App.

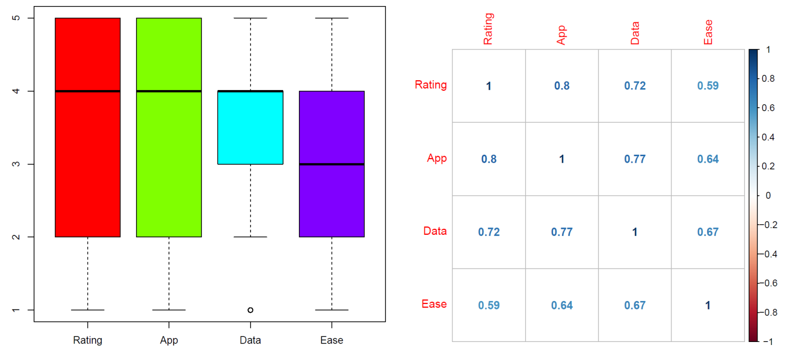


a) Students by Age Groups b) Ratings by Year

Fig. 3: Data Characteristics

The course project is used to let students have hands-on practice, and it is better for them to have practical experience in multiple aspects of data analytics. However, some Kaggle entries may provide a preprocessed data, which decreases the difficulty in data processing. Or, they may provide examples of research ideas or problems, which reduces the burdens of brainstorming or critical thinking. Therefore, we also ask instructors to give ratings in the two criteria "Data" and "Ease" for all the 70 items. These ratings basically reflect the degree of ease of the projects from the perspective of the instructors. And these ratings can be used to estimate the utility of the projects from the perspective of the instructors. Note the "App" refers to how students like the application or the domain of the data, while instructors have no limitations on it – that is the reason why we did not collect the instructor's rating on "App".

Students and instructors are the two stakeholders in this case study. The instructors respect students' choices and encourage them to look for any datasets they are interested in. They have no limitations on the criterion



a) Boxplot of Ratings b) Correlation Between Ratings

Fig. 4: Rating Characteristics

"App". However, the instructors encourage students to take advantage of this chance and select more challenging projects. The ease of the project is also considered in the final grading, because the final score depends on the degree of difficulty and their performance in the final projects. From the perspective of the students, some of them may prefer to select easier topics since they would like to save time and efforts so that can complete the projects easily and quickly. Some others may prefer to choose more challenging topics so that they can learn more by these hands-on practices. Therefore, MSRS is necessary to recommend appropriate projects to the students by balancing the needs of both students and instructors.

4 Utility-Based Multi-Stakeholder Recommendations

In this section, we introduce the utility-based multi-stakeholder recommendation models which were proposed in our previous work [58, 64, 60].

4.1 Utility-Based Multi-Stakeholder Framework

Utility-based recommendation is one of the recommendation models, according to the classification of RS by Burke [7]. A utility function is necessary to be built to capture the value of the item from the perspective of a user. The utility score associated with an item and a user can be used to rank the candidate items and produce the top-N recommendations to the end user.

The utility-based multi-stakeholder framework was first proposed in [63]. It is general enough for the multi-stakeholder recommendations.

- First of all, we need to figure out the stakeholders in a system. The utility function should be built to capture the value of the item from the perspective of each stakeholder. Each utility function can produce a utility score associated with an item and a stakeholder.

- The ranking score which are used to produce the top-N recommendations could be produced by an aggregation of the utility scores from the perspective of multiple stakeholders. The most straightforward function is a linear aggregation of these utility scores, where the weights in the linear function are the parameters to be learned in the recommendation process.
- Meantime, we need to define the objectives to balance the needs of multiple stakeholders and can utilize the multi-objective optimization methods to fine the optimal solution, as discussed in the Section 2.2.

4.2 Educational Multi-Stakeholder Recommendations

In this section, we introduce our previous work [64, 60] which build models for the educational case study by using the utility-based multi-stakeholder framework above. To better discuss our model, we list and explain the major notations in Table 2.

Table 2: Notations

Notation	**Explanations**
s	a student
p	a Professor or an instructor
t	an item
L	the top-N recommendation list
$U_{s,t}$	the utility of the item t from the perspective of student s
$U_{p,t}$	the utility of the item t from the perspective of instructor p
$U_{s,L}$	the utility of the list L from the perspective of student s
$U_{p,L}$	the utility of the list L from the perspective of instructor p
$R_{s,t}$	the multi-criteria rating vector given by student s on the item t
$R_{p,t}$	the multi-criteria rating vector given by instructor p on the item t
E_s	the expectation vector for the student s
E_p	the rating vector for the instructor p as minimal requirements

4.2.1 Utility Functions

Given a student s and a candidate item t, we first predict how s will rate t in the three criteria, "App", "Data" and "Ease", respectively by using the biased matrix factorization [26] which is a standard benchmark in traditional RS. These predicted multi-criteria ratings can be represented by a rating vector $R_{s,t}$. For each student, we assume there are student expectations which can be represented by the expected ratings in these three criteria. We denote it by the student expectation vector E_s. Namely, in E_s, we have a student's expectations on "App", "Data" and "Ease" as the underlying standards for the students to select the preferred projects. The values in E_s are unknown in our dataset and we will discuss how to learn them in the following paragraphs.

Note that the expectations are not always the "full-stack". The rating in each criterion is not the full (i.e., maximal rating value). Take the hotel reviews on the TripAdvisor.com for example, there may be several criteria, such as location, room size, cleanliness, and so forth. The ideal situation is that we would like to book a hotel with all five-star in these criteria. However, there are always some factors which may persuade us to lower our expectations, such as the budget for hotel bookings. Similarly, there are also factors in the case of educations, such as the capability of the students. It may result in the situation that not everyone would like to select more challenging projects in the class.

The utility of the item t from the perspective of the student s, can be denoted by $U_{s,t}$. It can be computed by using the similarity between the student's ratings on multiple criteria ($R_{s,t}$) and student expectations (E_s), as shown in Equation 1. The underlying assumption is straightforward – the user may like the item more if the item meets his or her expectations, i.e., the similarity value is higher.

$$U_{s,t} = Similarity(E_s, R_{s,t}) \tag{1}$$

In terms of the student expectations, we decide to learn E_s in advance by using the utility-based multi-criteria recommender (UBRec) [61], as instructed by our previous work [59]. UBRec is a traditional recommendation model which only considers the preferences of the end users. In UBRec, we use the $U_{s,t}$ as the score to rank the candidate items and produce the top-N item recommendations. Accordingly, we can use a learning-to-rank method to learn the student expectations E_s by maximizing the ranking metric normalized discounted cumulative gain (NDCG) [52].

The instructors encourage students to select more challenging course projects. However, there are always underperformed and outperformed students in the class, and instructors cannot require every student to select more challenging datasets. We can define E_s as the maximal expectations on the items from the perspective of the student. However, we cannot represent E_p as the unified expectation of the instructor since the instructor's expectations may vary from students to students. To simplify the problem, we set up a minimal requirement by the instructor p which can be described by the vector E_p associated with the criteria "Data" and "Ease". There are no requirements on the dimension "App", since students can select the projects in any domains or applications.

Recall that we have already collected the instructor's rating on each item in our data, and we use the $R_{p,t}$ to represent this rating vector. Therefore, the dissimilarity between E_p and $R_{p,t}$ can be used to denote the utility of an instructor, $U_{p,t}$, as shown in Equation 2. The reason why we use dissimilarity is because the E_p represents the minimal requirements, and the instructors would like the students to avoid much easier ones at least. The drawback of this dissimilarity function is that the model may suggest items with large dissimilarities regardless it is much easier or much more difficult. We alleviate this issue by proposing extended solutions in Section 4.3.

$$U_{p,t} = Dissimilarity(E_p, R_{p,t}) \tag{2}$$

Due to the fact that there is only one instructor in our data, we acquire the E_p from the instructor. It is 4 and 4 for the criteria “Data” and ”Ease” respectively. In other words, a project with rating 5 in “Data” and ”Ease” may not be suggested as the project in the class from the perspective of the instructor. We try to figure out the appropriate similarity measure to calculate the similarities in Equation 1 and Equation 2. Theoretically, any similarity or distance measures can be adopted. The best choice may vary from data to data. According to our previous experience [59], we found that the Pearson correlations and cosine similarity may not be reliable if the number of multiple criteria in the data is limited. In our experiments, we utilize the Euclidean distance. Assume there are two vectors m and n with same size k, Euclidean distance between m and n is the length of the line segment connecting them, as shown by Equation 3. More specifically, we compute the Euclidean distance between the rating and expectation vectors, normalize the distance to the scale [0, 1], and use 1 minus the normalized Euclidean distance to represent the similarity in our experiments.

$$distance(m, n) = \sqrt{\sum_{i=1}^{k} (m_i - n_i)^2} \tag{3}$$

4.2.2 Multi-Stakeholder Recommendations By Multi-Objective Learning

For each item to be recommended, we can calculate the utility score $U_{s,t}$ and $U_{p,t}$. We use a linear aggregation of these two scores to produce the ranking score which will be used to rank the items an generate the recommendation list L, as shown by Equation 4.

$$Ranking\ score(s, p, t) = \phi \times U_{s,t} + (1 - \phi) \times U_{p,t} \tag{4}$$

ϕ is the weight factor in scale [0, 1]. Note that 0.5 may not be the best choice for ϕ, since the distribution of $U_{s,t}$ and $U_{p,t}$ may be different. By using this linear aggregation, we rank the items by fusing the utilities of the item from the perspective of both students and instructors, which realizes the intrinsic goal of the multi-stakeholder recommendations.

As a result, whether we can balance the needs of multiple stakeholders depends on the value of ϕ. It can be learned through a process of multi-objective learning. As discussed in Section 2.2, there are three methods for multi-objective optimizations. In our work, we learn the optimal value of ϕ by using the Pareto optimizer. More specifically, we use the optimization algorithms built in the open-source library MOEA[2].

For the purpose of multi-objective learning, we can setup the learning objectives as follows:

[2] http://moeaframework.org/

- $U_{s,L}$ refers to the utility of the top-N recommendation list L from the perspective of the students s. It is the average of the $U_{s,t}$ over all items in L, while t is an item in the list L.
- $U_{p,L}$ refers to the utility of the top-N recommendation list L from the perspective of the instructor p. It is the average of the $U_{p,t}$ in L.
- The difference between $U_{s,L}$ and $U_{p,L}$, and we want to minimize this difference for the purpose of balance.
- The recommendation performance, such as precision, recall, NDCG. These metrics can be viewed as another representative of the student utilities. They may be decreased when we additionally consider the utility of instructors.

The multi-objective learning will minimize the difference between $U_{s,L}$ and $U_{p,L}$, and maximize other objectives. The optimal solution is expected to balance the needs of students and professors. It may decrease the recommendation performance since there could be conflicting interests in our case study. Namely, instructors would like students to select more challenging projects, while some students prefer to choose easier ones.

4.3 Extensions

There are two drawbacks in the multi-stakeholder recommendation models above. First, the similarity or distance calculations in the utility function may lead to the problem of over-/under-expectations. The issue refers to the situation that a user's rating on an item may lead to over-/under-expectations in comparison with the user's expectations on the items. Finally, it could result in false positives and/or false negatives in the recommendation process. Take Table 3 for example, the first three rows refer to user u's rating vectors on three items, while the last row refers to u's expectations to select a restaurant to dine in. It is clear that u's ratings on T_1 are under-expected, while his or her ratings on T_2 are over-expected. However, some of u's ratings on T_3 are under-expected, while others are over-expected. It results in the difficulty of deciding whether the user will like T_3.

Table 3: Example of Over-/Under-Expectation

User	**Item**	**Food**	**Service**	**Value**	**Ambiance**
u	T_1	2	2	2	2
u	T_2	4	4	4	4
u	T_3	1	4	2	1
u's expectation		3	3	3	3

The proposed method in Section 4.2 cannot tell whether the students would accept or reject an item if the estimated ratings falls in these situations. Our previous work [64] defined rules and used filters to remove candidate items to alleviate the problem of over-/under-expectations. An example of the rule could be "we should remove the item from the list of candidates if it is the case of over-expectation". Our results reveal that we can find better multi-stakeholder recommendation models if the appropriate rules can be applied. We also proposed to learn the penalties associated with the case of over-/under-expectations [62], while we do not need any domain knowledge to define the rules for the purpose of filtering.

Furthermore, students and instructors may have different perceptions on the degree of the difficulty of each project. A project which seems to be easy from the perspective of the instructor may be difficult for the students. Another work [59] proposed to adjust the students' or instructors' ratings on "Data" and "Ease" by a linear combination of the perceptions (i.e., ratings) by the students and instructors. Equation 5 gives an example, while we fuse $R_{s,t}$ and $R_{p,t}$ to update $R_{p,t}$, and $w1$ is the weight factor which will be additionally learned in the model. Similarly, the rating vector by instructors can be updated by Equation 6. Note that we only update the ratings on the dimension "Data" and "Ease" by using this linear aggregation. The utility function and the process of optimizations are the same as the ones discussed in Section 4.2. This method has been demonstrated to alleviate the problem and improve the models.

$$R_{s,t} = w_1 \times R_{s,t} + (1 - w_1) \times R_{p,t} \tag{5}$$

$$R_{p,t} = w_2 \times R_{p,t} + (1 - w_2) \times R_{s,t} \tag{6}$$

5 Offline and Online Experimental Design

In this section, we introduce and discuss our experimental design for offline and online evaluations, including baseline approaches, evaluation metrics, and user interfaces if necessary. Note that the offline experiments were completed in our previous work [64, 59], the major contribution of this paper is the exploration of online studies, especially the impact of explanation and transparency in MSRS.

5.1 Offline Experiments

5.1.1 Recommendation Models for Comparisons

In both offline and online experiments, we select four recommendation models for the purpose of comparisons:

– **alg1** refers to the utility-based multi-criteria recommender (UBRec) [61] which was demonstrated as the best baseline recommendation model. Note that this is a traditional recommendation model which was optimized by using the student preferences only.

– **alg2** denotes the Rank_p model which rank the items by considering the perspective of the instructors only. Simply, we rank the items by using the dissimilarity score between instructor expectations E_p and ratings $R_{p,t}$ to produce the top-N recommendations.
– **alg3** is the best utility-based multi-stakeholder recommendation model which additionally deals with the issue of over-/under-expectations, as described in our previous work [64]. The optimal value of ϕ is 0.9.
– **alg4** is the best utility-based multi-stakeholder recommendation model which additionally captures the perception of students and instructors by using preference corrections [59] as mentioned above. The optimal values for ϕ, w_1 and w_2 are 0.54, 0.98 and 0.42, respectively.

5.1.2 Evaluation Metrics

Recall that, we use the Pareto optimization to find the optimal parameters in our utility-based multi-stakeholder recommendation models. We develop a utility loss as the metric to select the optimal solution, while this loss can also be used as the offline evaluation metrics. More specifically, the method alg1 is the model which considers the student preferences only, while the method alg2 is the model which considers the instructor's standard only. The objectives produced by these two models can help us identify the maximal value in each objective. We use $\text{maxU}_{s,L}$, $\text{maxU}_{p,L}$, maxF_1 maxNDCG to denote these maximal objectives. As a result, the utility loss is average loss from three components – $\text{U}_{s,L}$, $\text{U}_{p,L}$, and recommendation performance, as shown by the Equation below. The model is better if the utility loss is smaller.

$$\begin{aligned} UtilityLoss = \frac{1}{3}(&\frac{maxU_{s,L} - U_{s,L}}{maxU_{s,L}} + \frac{maxU_{p,L} - U_{p,L}}{maxU_{p,L}} \\ &+ \frac{1}{2}(\frac{maxF_1 - F_1}{maxF_1} + \frac{maxNDCG - NDCG}{maxNDCG})) \end{aligned} \tag{7}$$

5.2 Online Experiments

We have two tasks in the online experiments – online evaluations and exploration of transparency on MSRS. The online evaluation is used to compare the recommendation models from the user experiences. In addition, we present explanations on the user interface to examine the impact of transparency.

5.2.1 Online Evaluations: The Rank of the Models

We built the models by using the four algorithms in Section 5.1.1, and used these models to produce the top-5 recommended projects for each student. These recommendations were stored in the MySQL database so that we can retrieve these data and present them on our Web survey system. The user interface used to evaluate the rank of these four models can be shown by Figure 5 which is the 1st page when the students entered our Web survey system. On this page, we display the four list of top-5 recommendations which were produced by the four recommendation models as mentioned above. The sequence of the four lists is randomly produced and blended for each student. The textual descriptions on this page asks the students to "evaluate the quality of the four list of recommendations and give rank 1 (the best) to 4 (the worst) to them". The ranks must be unique and cannot be duplicated. Students can click the URL of each recommended project to visit the corresponding Kaggle webpage to learn more details of each recommended project. Once we collected the inputs from the students, we can use ANOVA or hypothesis testing to examine whether there is a significant difference in the rank of these recommendation models. The findings will be compared with the results in the offline experiments to discover more insights.

TOP	TITLE
1	French presidential election
2	Circadian Rhythm in the Brain
3	European Soccer Database
4	Mushroom Classification
5	Emotion; Aging; and Sentiment Over Time

Rank:

TOP	TITLE
1	European Soccer Database
2	French presidential election
3	Circadian Rhythm in the Brain
4	Lending Club Loan Data
5	Pima Indians Diabetes Database

Rank:

TOP	TITLE
1	European Soccer Database
2	Circadian Rhythm in the Brain
3	French presidential election
4	Lending Club Loan Data
5	Speed Dating Experiment

Rank:

TOP	TITLE
1	European Soccer Database
2	Lending Club Loan Data
3	Pima Indians Diabetes Database
4	Iris Species
5	NBA Free Throws

Rank:

Fig. 5: User Interface for the Rank of the Algorithms

5.2.2 Transparency of ϕ in Multi-Stakeholder Recommendation Models

We developed two Web pages to examine the impact of transparency in alg3, and another two pages for alg4. The pages for alg3 are similar to the pages for alg4. First, I used the user interfaces for alg3 as an example in the following introductions and discuss the differences in the pages for alg4. Note that there is a weight factor ϕ in both alg3 and alg4. We built different models by varying the value of α from 0 to 1 with 0.1 in each step and produced the top5 recommendations for each student which were stored in the MySQL database. As we mentioned before, there are two existing methods to enhance the transparency in the recommendation models – one is presenting explanations on the user interface, while another one is providing user controls in the interactive user interfaces. In our work, we propose a new way – disclosing and explaining the parameter ϕ in our MSRS model, providing user controls to select the optimal value, and finally comparing the selected values between with and without explanations. The design of the interactive user interfaces can be introduced as follows.

The two pages for alg3 are similar, both have a same interactive user interface as shown by Figure 6. The interactive user interface provides a user control by using the slider bar. On the first page, the students were told that there was a parameter α (i.e., same as ϕ in alg3) scales in [0, 1] in the recommendation system, and asked students to adjust the value of α by using the slider and selected the optimal value of *alpha* when they observe the best list of recommendations. By moving the slider, the value of α will be increased or decreased by 0.1, and the web system will automatically retrieval and display the list of recommendations associated with the α which were stored in the MySQL database. Note that, we did not explain the notion of *alpha* here, and students have no idea about how the recommendation model works and how the α is associated with the model.

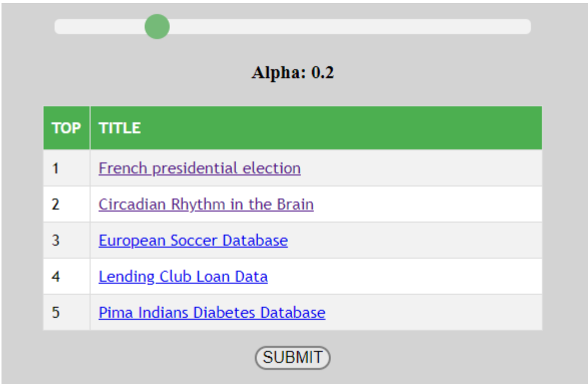


Fig. 6: UI for the Transparency of ϕ

Once the student selected the optimal *alpha* on this page, they will be directed to the next page which shows a same interactive user interface in Figure 6. By contrast, we changed the texts above the solder bar – we explained what does α mean, and let the students select the optimal value of *alpha* after our explanations. Our textual explanations on this page are, "*The value of α indicates the degree of your own satisfaction with the recommendations, while 1 - α can tell the satisfaction of the professor. Note that the recommended projects may be much more difficult if you reduce the value of α. Please adjust the value of α and figure out which list of recommendations is best for you*". By this way, we design and incorporate two elements in the transparency – we explain the role of α, and we let the students to select the optimal value of α. The experiments were designed in this way for two purposes – on one hand, we would like to observe the different choices of α with versus without explaining α; on the other hand, we would like to compare the optimal choice of α by using offline experiments and the online user studies.

We have two additional pages designed for ϕ in alg4. The two pages are almost the same with the pages for alg3. The only difference is that we told student there was a parameter β instead of α to let students avoid confusions and believe that the models are different, and the two parameters may take different roles. In fact, both α in alg3 and β in agl4 refer to the same ϕ in the multi-stakeholder recommendations as shown in Equation 4.

5.3 Response Collections

Recall that there are 269 students in our data for the purpose of offline experiments, we distribute this Web survey to all these students to collect their inputs for online studies. However, some students have already graduated, and their school emails may not work anymore. Finally, we received valid and complete responses from 52 students. Furthermore, instructor is another stakeholder in our case study. We also let the instructor fill the Web survey again on behalf of these 52 students. Namely, the instructor needs to use email address of these 52 students and fill in the questionnaire, but the instructor gave the inputs from the perspective of instructors. In this case, we can evaluate the rank of the algorithms and the impact of transparency from the perspective of both students and instructors.

Table 4: Example of Responses

UserID	α_s	α_s^e	β_s	β_s^e	**alg1**	**alg2**	**alg3**	**alg4**
1002	0.4	0.6	0.7	0.8	2	1	3	4
1005	0.7	0.1	0.7	0.1	2	3	1	4
1008	0.3	0.4	0.6	0.9	1	3	2	4
1020	0	0	0.9	0.9	4	1	3	2
...	...	...	...	...	...	...	...	...

The example of responses can be shown by Table 4. Note that we have two such matrices – one is the responses from students, while another one has the responses from the perspective of instructors.

The values in the columns alg1, alg2, alg3, and alg4 refer to the ranks of each algorithm. We use α_s and β_s to denote the optimal α and β selected by students without explanations, while α_s^e and β_s^e are the optimal values for α and β selected by $s$tudents after the explanations. Accordingly, we use α_p and β_p, α_p^e and β_p^e to indicate the optimal value selected by the instructors or professors.

6 Experimental Results and Findings

In this section, we present and discuss our results and findings in both offline and online experiments.

6.1 Offline Experimental Results

We use 5-fold cross validation for the purpose of offline evaluations, and Table 5 presents the offline experimental results. We present the utility of top-5 recommendation list L from the perspective of students $U_{s,L}$ and instructors $U_{p,L}$ respectively, as well as the recommendation performance based on the F-1 measure (i.e., F_1) and NDCG. The values in bold are the best values in each objective by the baseline approaches.

Table 5: Results of the Offline Experiments

	$\mathrm{U}_{s,L}$	$\mathrm{U}_{p,L}$	F_1	NDCG	Utility Loss	ϕ
UBRec (alg1)	**0.1812**	0.1338	**0.0845**	**0.1264**	0.1837	N/A
Rankp (alg2)	0.072	**0.2982**	0.0274	0.039	0.4287	N/A
alg3	0.1816	0.1947	0.0719	0.1044	0.1688	0.9
alg4	0.2029	0.1418	0.0838	0.1262	0.1366	0.54

Based on the results above, we can find out that the alg4 which applies the preference corrections is the best model, while both the alg3 and alg4 (i.e., the two multi-stakeholder recommendation models) are better than the baseline models. The optimal ϕ in alg3 and alg4 are 0.9 and 0.54, respectively. Moreover, by comparing the results by baselines and MSRS models, we can observe that the benefits of instructors (i.e., $\mathrm{U}_{p,L}$) were improved at the cost of recommendation performance.

One of the weaknesses in the offline experiments is that the number of recommended items is limited. We only evaluate the models based on top-5 recommendations, since we would like to keep a consistency between the offline and online experiments. We would like to present top-5 recommended items to the students in the online user studies, to reduce user efforts in examining a longer recommendation list. However, it may also decrease the reliability of the results.

6.2 Online Experimental Results

6.2.1 The Rank of the Models

First of all, we present the rank of the algorithms from the perspective of students and instructors, as shown by Figure 7. We also present the results of ANOVA and paired t-test in the Table 6 and 7.

From the perspective of the students, there are no significant differences among these four algorithms, where the p-value in ANOVA is 0.384. By contrast, there are some significant differences from the perspective of instructors, while the only exception is the alg2 and alg4 (p-value = 0.11) if we use 95% as the confidence level.

Table 6: Student's Perspective[a]

	alg1	alg2	alg3
alg2	0.45		
alg3	0.40	1	
alg4	0.12	0.45	0.47

[a] ANOVA, p-value = 0.384

Table 7: Instructor's Perspective [b]

	alg1	alg2	alg3
alg2	1.22e-18		
alg3	1.14e-10	1.58e-3	
alg4	4.18e-13	0.11	0.05

[b] ANOVA, p-value < 2e-16

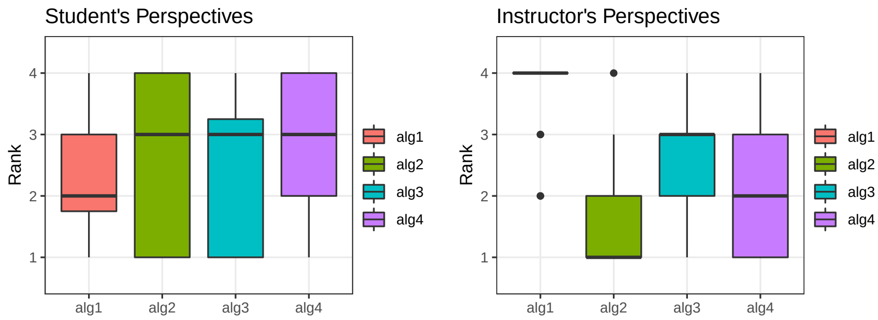


Fig. 7: Results: The Rank of the Algorithms

The descriptive statistics of the rank of the algorithms can be viewed from Table 8. Particularly, alg1 which produces recommendations by considering student preferences only is considered as the worst one from the perspective of instructors, which can be observed from the statistics in the Table 8 and the boxplot in Figure 7. However, there are no significant differences in alg2 which produce recommendations by considering the perspective of instructors only and alg4 which uses preference corrections to capture different perceptions of the students and instructors.

Table 8: Descriptive Statistics of the Algorithm Ranks

		alg1	alg2	alg3	alg4
Student's Perspective	mean	2.31	2.5	2.5	2.69
	sd	1.02	1.18	1.15	1.13
Instructor's Perspective	mean	3.76	1.73	2.45	2.06
	sd	0.59	0.87	0.88	0.88

Recall that maximizing the preference of one stakeholder may hurt the utility of the item in view of other stakeholders. Particularly, in our case, instructors encourage students to select more challenging projects, while some students may prefer to select easier ones. In this view, the results of the algorithm from the perspective of students are still positive, even if there are no significant differences. It is because students may not feel hurt, because there are no significant differences between the MSRS models and the alg1 model which produces the recommendations by considering student preferences only. It may tell that the recommendations by multi-stakeholder models are still tolerant to the students. However, it is surprising to see students did not observe the differences in alg1 and alg2. One of the potential reasons is that there are no big differences in the top-5 recommendations. Students may observe differences if we gave them more recommended items.

6.2.2 *The Impact of Enhanced Transparency*

The optimal selected α and β values can be observed from Figure 8. The parameters with subscript s and p refer to the values selected by students and instructors respectively, while the parameters with the superscript e refer to the selected values after the explanations.

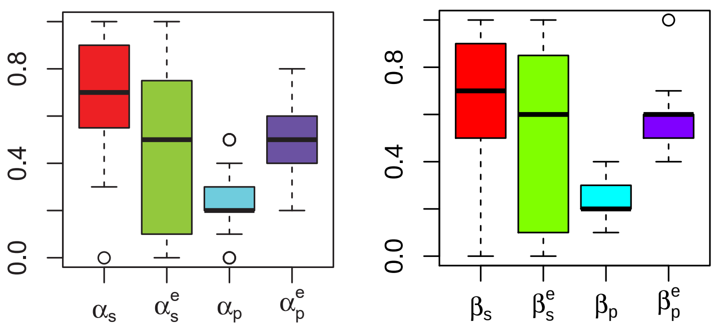


Fig. 8: Results of the Optimal α and β

First, we compare a_s versus a_s^e and b_s versus b_s^e, we can observe that the optimal value of α and β will be decreased after the explanations from the perspective of students, while the difference is significant with p-values $<$ 2.0e-12. By contrast, from the perspective of instructors, the a_s^e and b_s^e will be increased after the explanations. In other words, both students and instructors prefer to make certain concessions to balance the needs of each other. More descriptive statistics can be described by Table 9 and 10.

In addition, we compare a_s and a_p, a_s^e and a_p^e, as well as b_s and b_p, b_s^e and b_p^e. The results are not surprising. There are significant differences between α and β from the perspective of students and instructors without explanations.

Table 9: Student's Perspective

	mean	sd	se
a_s	0.7	0.23	0.03
a_s^e	0.46	0.36	0.05
b_s	0.68	0.22	0.03
b_s^e	0.49	0.36	0.05

Table 10: Instructor's Perspective

	mean	sd	se
a_p	0.22	0.12	0.02
a_p^e	0.5	0.13	0.02
b_p	0.21	0.09	0.01
b_p^e	0.53	0.1	0.01

However, after explanations, the differences between a_s^e and a_p^e, and b_s^e and b_p^e are not significant. Interestingly, after the explanations, both students and instructors prefer to set the optimal α and β values as close to 0.5. Particularly, the *beta* value in alg4 will be close to 0.5, since the preference corrections will help capture the different perceptions of students and instructors, which balances the distributions of item utilities in view of students and instructors.

Finally, we compare the optimal α and β values in the offline experiments and online user studies. The values for ϕ or β in alg4 are actually consistent in both offline and online results, if we explained this parameter to the end users. However, the values for ϕ or α in alg3 are very different. It is 0.90 in the offline experiments, but 0.46 and 0.5 by the students and instructors in the online user studies. One possible explanation is that students or instructors tend to avoid extreme values after the explanations of the parameters, since they would like to balance the needs of two stakeholders.

One of the weaknesses in this study is that the number of instructors is limited. It is able to obtain more reliable results or patterns if we have multiple instructors. However, it is difficult to collect the data with multiple instructors at the current stage.

6.3 Summary of Findings

In this paper, we made the first attempt to evaluate multi-stakeholder recommendations by using both offline experiments and user studies. Particularly, we build an interactive user interfaces which let the users tune up the key parameters in the recommendation model. With extra explanations of these key parameters, we try to observe how different stakeholders react on the selections for these parameters, to examine the effect of transparency. Our findings can be summarized as follows.

– In the offline experiments, alg3 and alg4 outperform the baseline approaches, and alg4 gives the best model. By contrast, in the online user studies, students did not present significant differences on the recommendation lists produced by different algorithms, while the instructor prefers the recommendation list produced by alg2 or alg4. We consider this result as positive ones since the students are tolerant to the multi-stakeholder recommendations. As mentioned before, the recommendations may hurt the satisfaction of the end users, but actually students did not feel that because there are no significant differences among these algorithms in online studies.

– There are significant differences in the optimal choice of the weight factor ϕ in comparison of with versus without explaining the parameter to the subjects. Students would like to lower their weights, while instructors can also make certain concessions. The enhanced transparency by the explanation of the parameter ϕ and the user controls allows the stakeholders to balance the needs of others and select the list of recommendations that they are tolerant with.
– In terms of the optimal weight factor ϕ in the offline experiments versus online user studies, we found that the values are consistent in the alg4, while they are very different in the alg3. Note that there are other offline solutions in the Pareto optimal set which deliver a similar utility loss, while we selected the solution with smallest loss. The value of ϕ may be different in these solutions, even if the utility loss is similar. Another possible reason could be that we select the best offline models based on the data associated with 269 students, while the results in online studies were produced based on the 52 subjects only.

6.4 Limitations and Lessons Learned

We believe that there are at least three limitations or concerns which are worth pointing out.

– *The number of instructors is limited.* We have two stakeholders in the educational study – students and instructors. We have multiple students in the data, while we only have one instructor. In the general context of multi-stakeholder recommendations, there may be multiple individuals who belong to a type of the stakeholder. We would like to point out that the special characteristics in this case study leads to such a limitation. If there are multiple instructors teaching a same course, instructors may have different expectations or requirements, which results in the general context of MSRS. However, there is only one instructor who teaches the data analytics course in our department. Alternatively, we may bring more courses in the case studies, which can increase the number of courses and instructors, but different courses may define different projects or items to be recommended. In other words, the project recommendations are usually course-dependent, and it could be even instructor-dependent, if there is only a single instructor for each individual course. We believe our experimental design and findings are still worth learning, especially for the applications in educations. The experiments could be more complicated if it falls in the general context of multi-stakeholder recommendations.
– *There are 269 students in the offline experiments, while there are only 52 subjects in the online studies.* Actually, we collected the dataset over several years, and evaluated our models in the offline experiments. Later, we realize the importance of the online studies and started to work for online experiments. It is possible that the results may vary if we examine the models for different groups of the students. However, the results based

on the 52 subjects in the offline experiments may not be reliable, since the ratings associated with these 52 subjects are too limited.

– *The number of recommended items is limited.* In our online studies, we present the top-5 recommended items to the subjects. We did not observe the significance differences among the four recommendation models from the perspective of the students. One of the possible reasons is the limited number of recommended items. If you show a longer list of recommendations to the students, they may be able to response differently. However, it also brings more burdens or human efforts in the user studies. Therefore, there are always pros and cons, if we use a shorter or longer list of recommendations.

7 Conclusions and Future Work

In contrast to traditional RS, MSRS try to produce the item recommendations and balance the needs of different stakeholders by taking multiple stakeholders into considerations. Existing work on MSRS only focused on offline experiments, while we believe the user experiences and subjective feedbacks from different stakeholders are very important. In this paper, we make the first attempt to exploit the online user studies for MSRS by using an educational dataset as the case study. Particularly, we enhance the transparency by designing specific user interfaces which provide explanations of the key parameters in MSRS and interactive user controls to let the subjects select the optimal value for the parameters. We observe that the explanation of the parameters enhances the transparency because the users would like to make concessions in the parameter selections by considering the needs of other stakeholders.

MSRS is novel and promising, there are plenty of the research topics we can work on in the future. For example, we may perform online studies among different multi-stakeholder recommendation solutions, rather than the solutions based on the Pareto optimization only in this paper. In addition, we explain the weight factor in our online studies, while we can also examine the impact by the explanations of the user expectations. Furthermore, we would like to exploit the research on the applications with more than two types of the stakeholders, in order to discover more insights in this area.

References

1. Abdollahpouri, H., Adomavicius, G., Burke, R., Guy, I., Jannach, D., Kamishima, T., Krasnodebski, J., Pizzato, L.: Multistakeholder recommendation: Survey and research directions. User Modeling and User-Adapted Interaction **30**(1), 127–158 (2020)
2. Adomavicius, G., Kwon, Y.: New recommendation techniques for multicriteria rating systems. IEEE Intelligent Systems **22**(3), 48–55 (2007)
3. Akoglu, L., Faloutsos, C.: Valuepick: Towards a value-oriented dual-goal recommender system. In: 2010 IEEE International Conference on Data Mining Workshops, pp. 1151–1158. IEEE (2010)
4. Al-Otaibi, S.T., Ykhlef, M.: A survey of job recommender systems. International Journal of the Physical Sciences **7**(29), 5127–5142 (2012)

5. Anwar, S., McMillan, R., Zheng, M.: Bidding behavior in competing auctions: Evidence from ebay. European Economic Review **50**(2), 307–322 (2006)
6. Brozovsky, L., Petricek, V.: Recommender system for online dating service. arXiv preprint cs/0703042 (2007)
7. Burke, R.: Hybrid recommender systems: Survey and experiments. User modeling and user-adapted interaction **12**(4), 331–370 (2002)
8. Burke, R., Abdollahpouri, H.: Educational recommendation with multiple stakeholders. In: 2016 IEEE/WIC/ACM International Conference on Web Intelligence Workshops (WIW), pp. 62–63. IEEE (2016)
9. Burke, R.D., Abdollahpouri, H., Mobasher, B., Gupta, T.: Towards multi-stakeholder utility evaluation of recommender systems. In: UMAP (Extended Proceedings) (2016)
10. Chen, J., Dong, H., Wang, X., Feng, F., Wang, M., He, X.: Bias and debias in recommender system: A survey and future directions. arXiv preprint arXiv:2010.03240 (2020)
11. Chen, L.S., Hsu, F.H., Chen, M.C., Hsu, Y.C.: Developing recommender systems with the consideration of product profitability for sellers. Information Sciences **178**(4), 1032–1048 (2008)
12. Deldjoo, Y., Anelli, V.W., Zamani, H., Bellogin, A., Di Noia, T.: A flexible framework for evaluating user and item fairness in recommender systems. User Modeling and User-Adapted Interaction (2021)
13. Deng, F.: Utility-based recommender systems using implicit utility and genetic algorithm. In: 2015 International Conference on Mechatronics, Electronic, Industrial and Control Engineering. Atlantis Press (2015)
14. Deuze, M.: What is journalism? professional identity and ideology of journalists reconsidered. Journalism **6**(4), 442–464 (2005)
15. Diakopoulos, N., Koliska, M.: Algorithmic transparency in the news media. Digital Journalism **5**(7), 809–828 (2017)
16. Ekstrand, M.D., Azpiazu, I.M., Wright, K.L., Pera, M.S.: Retrieving and recommending for the classroom. ComplexRec **6**(2018), 14 (2018)
17. Evans, J.R., Mathur, A.: The value of online surveys. Internet research (2005)
18. Felfernig, A., Gula, B.: Consumer behavior in the interaction with knowledge-based recommender applications. na (2006)
19. Flyverbom, M.: Digital age— transparency: Mediation and the management of visibilities. International Journal of Communication **10**, 13 (2016)
20. Fricker, R.D., Schonlau, M.: Advantages and disadvantages of internet research surveys: Evidence from the literature. Field methods **14**(4), 347–367 (2002)
21. Guttman, R.H.: Merchant differentiation through integrative negotiation in agent-mediated electronic commerce. Ph.D. thesis, Massachusetts Institute of Technology (1998)
22. Herlocker, J.L., Konstan, J.A., Riedl, J.: Explaining collaborative filtering recommendations. In: Proceedings of the 2000 ACM conference on Computer supported cooperative work, pp. 241–250 (2000)
23. Jannach, D., Karakaya, Z., Gedikli, F.: Accuracy improvements for multi-criteria recommender systems. In: Proceedings of the 13th ACM conference on electronic commerce, pp. 674–689 (2012)
24. Johnson, H., Johnson, P.: Explanation facilities and interactive systems. In: Proceedings of the 1st international conference on Intelligent user interfaces, pp. 159–166 (1993)
25. Kauffman, R.J., Wood, C.A.: Running up the bid: Modeling seller opportunism in internet auctions. AMCIS 2000 Proceedings p. 376 (2000)
26. Koren, Y., Bell, R., Volinsky, C.: Matrix factorization techniques for recommender systems. Computer **42**(8) (2009)
27. Krebs, L.M., Alvarado Rodriguez, O.L., Dewitte, P., Ausloos, J., Geerts, D., Naudts, L., Verbert, K.: Tell me what you know: Gdpr implications on designing transparency and accountability for news recommender systems. In: Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems, p. LBW2610. ACM (2019)
28. Kwak, M.: Web sites learn to make smarter suggestions. MIT Sloan Management Review **42**(4), 17–17 (2001)

29. Lee, M.K., Kusbit, D., Metsky, E., Dabbish, L.: Working with machines: The impact of algorithmic and data-driven management on human workers. In: Proceedings of the 33rd annual ACM conference on human factors in computing systems, pp. 1603–1612 (2015)
30. Li, Z., Fang, X., Bai, X., Sheng, O.R.L.: Utility-based link recommendation for online social networks. Management Science **63**(6), 1938–1952 (2017)
31. Lin, F.R., Huang, S.L., Yang, Y.C.: Using radial basis function networks to model multi-attribute utility functions. In: Proceedings of the 4th Workshop on e-Business (2005)
32. Lin, X., Chen, H., Pei, C., Sun, F., Xiao, X., Sun, H., Zhang, Y., Ou, W., Jiang, P.: A pareto-efficient algorithm for multiple objective optimization in e-commerce recommendation. In: Proceedings of the 13th ACM Conference on Recommender Systems, pp. 20–28 (2019)
33. Louca, R., Bhattacharya, M., Hu, D., Hong, L.: Joint optimization of profit and relevance for recommendation systems in e-commerce. In: Workshop on Recommendation in Multi-stakeholder Environments, in conjunction with the 13th ACM Conference on Recommender Systems, vol. 19, p. 2019 (2019)
34. Lyons, J.B.: Being transparent about transparency: A model for human-robot interaction. In: 2013 AAAI Spring Symposium Series (2013)
35. Lyons, J.B., Sadler, G.G., Koltai, K., Battiste, H., Ho, N.T., Hoffmann, L.C., Smith, D., Johnson, W., Shively, R.: Shaping trust through transparent design: theoretical and experimental guidelines. In: Advances in human factors in robots and unmanned systems, pp. 127–136. Springer (2017)
36. McSherry, D.: Explanation in recommender systems. Artificial Intelligence Review **24**(2), 179–197 (2005)
37. Musto, C., Narducci, F., Lops, P., de Gemmis, M., Semeraro, G.: Linked open data-based explanations for transparent recommender systems. International Journal of Human-Computer Studies **121**, 93–107 (2019)
38. Nguyen, P., Dines, J., Krasnodebski, J.: A multi-objective learning to re-rank approach to optimize online marketplaces for multiple stakeholders. arXiv preprint arXiv:1708.00651 (2017)
39. Nilashi, M., Jannach, D., bin Ibrahim, O., Esfahani, M.D., Ahmadi, H.: Recommendation quality, transparency, and website quality for trust-building in recommendation agents. Electronic Commerce Research and Applications **19**, 70–84 (2016)
40. O'Donovan, J., Smyth, B.: Trust in recommender systems. In: Proceedings of the 10th international conference on Intelligent user interfaces, pp. 167–174 (2005)
41. Pizzato, L., Rej, T., Chung, T., Koprinska, I., Kay, J.: Recon: a reciprocal recommender for online dating. In: Proceedings of the fourth ACM conference on Recommender systems, pp. 207–214. ACM (2010)
42. Pu, P., Chen, L., Hu, R.: A user-centric evaluation framework for recommender systems. In: Proceedings of the fifth ACM conference on Recommender systems, pp. 157–164. ACM (2011)
43. Rader, E., Cotter, K., Cho, J.: Explanations as mechanisms for supporting algorithmic transparency. In: Proceedings of the 2018 CHI conference on human factors in computing systems, pp. 1–13 (2018)
44. Rekimoto, J., Green, M.: The information cube: Using transparency in 3d information visualization. In: Proceedings of the Third Annual Workshop on Information Technologies & Systems, pp. 125–132 (1993)
45. Shin, D., Park, Y.J.: Role of fairness, accountability, and transparency in algorithmic affordance. Computers in Human Behavior **98**, 277–284 (2019)
46. Shortliffe, E.H., Buchanan, B.G.: Rule-based expert systems: the MYCIN experiments of the Stanford Heuristic Programming Project. Addison-Wesley Publishing Company (1985)
47. Sinha, R., Swearingen, K.: The role of transparency in recommender systems. In: CHI'02 extended abstracts on Human factors in computing systems, pp. 830–831. ACM (2002)
48. Sürer, Ö., Burke, R., Malthouse, E.C.: Multistakeholder recommendation with provider constraints. In: Proceedings of the 12th ACM Conference on Recommender Systems, pp. 54–62 (2018)
49. Thomaz, A.L., Breazeal, C.: Transparency and socially guided machine learning. In: 5th Intl. Conf. on Development and Learning (ICDL) (2006)

50. Tintarev, N., Masthoff, J.: A survey of explanations in recommender systems. In: 2007 IEEE 23rd international conference on data engineering workshop, pp. 801–810. IEEE (2007)
51. Tsai, C.H., Brusilovsky, P.: Providing control and transparency in a social recommender system for academic conferences. In: Proceedings of the 25th Conference on User Modeling, Adaptation and Personalization, pp. 313–317 (2017)
52. Valizadegan, H., Jin, R., Zhang, R., Mao, J.: Learning to rank by optimizing ndcg measure. In: Advances in neural information processing systems, pp. 1883–1891 (2009)
53. Wang, H.F., Wu, C.T.: A mathematical model for product selection strategies in a recommender system. Expert Systems with Applications **36**(3), 7299–7308 (2009)
54. Wright, K.B.: Researching internet-based populations: Advantages and disadvantages of online survey research, online questionnaire authoring software packages, and web survey services. Journal of computer-mediated communication **10**(3), JCMC1034 (2005)
55. Yu, H., Liu, C., ZHANG, F.: Reciprocal recommendation algorithm for the field of recruitment. Journal of Information and Computational Science **8**(16), 4061–4068 (2011)
56. Zarsky, T.: The trouble with algorithmic decisions: An analytic road map to examine efficiency and fairness in automated and opaque decision making. Science, Technology, & Human Values **41**(1), 118–132 (2016)
57. Zheng, Y.: Personality-aware decision making in educational learning. In: Proceedings of the 23rd International Conference on Intelligent User Interfaces Companion, p. 58. ACM (2018)
58. Zheng, Y.: Utility-based multi-stakeholder recommendations by multi-objective optimization. In: Proceedings of the 2018 IEEE/WIC/ACM International Conference on Web Intelligence. IEEE (2018)
59. Zheng, Y.: Multi-stakeholder personalized learning with preference corrections. In: Proceedings of the 18th IEEE International Conference on Advanced Learning Technologies (ICALT). IEEE (2019)
60. Zheng, Y.: Preference corrections: capturing student and instructor perceptions in educational recommendations. Smart Learning Environments **6**(1), 1–15 (2019)
61. Zheng, Y.: Utility-based multi-criteria recommender systems. In: Proceedings of the ACM Symposium on Applied Computing. ACM (2019)
62. Zheng, Y.: Penalty-enhanced utility-based multi-criteria recommendations. Information **11**(12), 551 (2020)
63. Zheng, Y., Dave, T., Mishra, N., Kumar, H.: Fairness in reciprocal recommendations: A speed-dating study. In: Adjunct Publication of the 26th Conference on User Modeling, Adaptation and Personalization, pp. 29–34. ACM (2018)
64. Zheng, Y., Ghane, N., Sabouri, M.: Personalized educational learning with multi-stakeholder optimizations. In: Adjunct Proceedings of the ACM conference on User Modelling, Adaptation and Personalization. ACM (2019)
65. Zihayat, M., Ayanso, A., Zhao, X., Davoudi, H., An, A.: A utility-based news recommendation system. Decision Support Systems **117**, 14–27 (2019)

Yong Zheng is an Assistant Professor at Department of Information Technology and Management, College of Computing, Illinois Institute of Technology, USA. He received a Ph.D. degree in Computer and Information Sciences from DePaul University, USA. His research interests lie in user modeling and recommender systems. He joined the organization committee and helped organize related academic conferences in these areas, including ACM RecSys, ACM UMAP, ACM IUI conferences, etc.

Juan Ruiz Toribio is currently a Business Analyst at Kearney, Spain. He obtained the Master's degree in Information Technology and Management from Illinois Institute of Technology, USA and Universidad Politécnica de Madrid, Spain. He worked on the topic of this paper during his studies at Illinois Institute of Technology, USA in 2019.