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

Conflict resolution in group decision making: insights from a simulation study

Thuy Ngoc Nguyen · Francesco Ricci · Amra Delic · Derek Bridge

Received: date / Accepted: date

Abstract An individual's conflict resolution styles can have a large impact on the decision making process of a group. This impact is affected by a variety of factors, such as the group size, the similarity of the group members, and the type of support offered by the recommender system, if the group is using one. Measuring the effect of these factors goes beyond the capability of a live user study. In this article we show that simulation-based experiments can be effectively exploited to analyse the effect of the group members' conflict resolution styles and to help researchers to formulate additional research hypotheses, which could be individually tested in ad hoc user studies. We therefore propose a group discussion procedure that simulates users' actions while trying to make a group decision. The simulated users adopt alternative conflict resolution styles derived from the Thomas-Kilmann Conflict Model. The simulation procedure is informed by the analysis of real users' interaction logs with a group discussion support system.

Our experiments are conducted on scenarios characterized by four group factors, namely, conflict resolution style, inner-group similarity, interaction length and group size. We demonstrate the effect of these factors on the recommendation quality. This is measured by the loss in the utility obtained by an

Thuy Ngoc Nguyen
Free University of Bozen-Bolzano, Bolzano, Italy,
E-mail: ngoc.nguyen@unibz.it

Francesco Ricci
Free University of Bozen-Bolzano, Bolzano, Italy,
E-mail: fricci@unibz.it

Amra Delic
TU Wien, Vienna, Austria,
E-mail: amra.delic@tuwien.ac.at

Derek Bridge
Insight Centre for Data Analytics, University College Cork, Cork, Ireland,
E-mail: derek.bridge@insight-centre.org

individual when choosing the recommended group choice rather than his/her individual best choice. We also measure the difference between the highest and lowest utility that the group members obtain, in order to understand the fairness of the group recommendation identified by the system.

The experimental results show (among other findings) that if group members have similar tastes then groups composed of users with the competing conflict resolution style obtain the largest utility loss, compared to groups whose members adopt the cooperative styles (accommodating and collaborating), and yet, whatever their conflict resolution styles, there is no distinct difference in their utility for the group choice (they are treated equally). Conversely, when group members have diverse preferences, the average utility loss of competing members is still the largest, but the differences in their utility is the lowest (they all get a similar but lower utility). Some of the findings of our simulation experiments also match observations made in real group discussions and they pave the way for new user studies aimed at further supporting the reported findings.

Keywords Group decision-making process · Conflict resolution styles · Group recommendations · Simulation design

1 Introduction

Recommender Systems (RSs) are tools that support individuals to make decisions by suggesting items that are likely to meet their needs and interests [37]. In many realistic scenarios the subject is looking for items to be consumed not individually but in a group. For example, a group of friends may be looking for a destination to visit together, so the recommendations are expected to satisfy all group members. These situations have brought up new challenges and led to the research in Group Recommender Systems (GRSs) [20, 27].

GRSs must focus on group decision-making processes, which, by their very nature, embrace conflict as an unavoidable consequence since different group members may have diverse outlooks and preferences, which could pull them in contrasting directions. In conflict situations, these individuals could either take into account the interests of the others or outright ignore them. This kind of behaviour, as pointed out by the Thomas-Kilmann Conflict Mode Instrument (TKI) [21, 41], can be characterized by two fundamental dimensions, namely assertiveness and cooperativeness, that are the extent to which an individual attempts to satisfy her own and other people's preferences, respectively. These two dimensions of user behaviour can be used to identify four possible conflict resolution styles: *accommodating*, *competing*, *avoiding*, and *collaborating*.

Previous research has used TKI in group recommendations to weigh the influence of group members; a person who is more assertive is assumed to have greater influence and hence is given a greater weight in the preference aggregation step of the GRS [35, 36]. However, the impact of the users' conflict resolution styles on the outcome of the decision making process supported by a GRS has not yet been analysed. Moreover, the effect of the group members' conflict resolution styles could be interconnected with other factors. For instance, it is not clear, whether the length of group interactions or the group size can mediate the effect of the group members' conflict resolution styles.

In this work, we therefore carry out an extensive analysis of the combined effect of four factors affecting the decision making of a group that uses a recommender system: *conflict resolution style*, *inner-group similarity*, *length of group interaction* and *group size*. We note that studying the impact of the above-mentioned four group factors on the quality of group recommendations is unlikely to be done by the analysis of real users due to the large variety of possible group situations that would need to be extensively evaluated. Thus, we have designed a simulation procedure where the dynamics of individuals' actions in alternative group discussion situations characterized by the four group factors is modelled. The proposed flow and structure of the simulated group discussion is informed by observations of real users' actions while interacting with a GRS that we have previously implemented and evaluated. The goal is to retain as much as possible the essential features of a realistic system-mediated group discussion.

The quality of the simulated recommendations and group choices is measured by observing the *Mean Individual Loss* — MIL, i.e., the average of the differences between each group member's utility of the collective choice and the utility of the item that each individual could have chosen if he had only optimized his personal choice. This metric is motivated by the principle of process losses and gains in social psychology, which focuses on how a group setting affects individual performance by comparing the performance of group members when they work in groups with what they would have achieved if they had worked on their own [16]. Moreover, due to conflicting situations, the outcome of a group discussion can bring winning and losing experiences, signifying that the recommendation that is the best for one user may be the worst for another. Therefore, we evaluate the quality of group recommendations by considering the *Max Min Difference* — MMD, i.e., the difference in utility between the group's winner and loser. The winners and losers are defined as those who receive the maximum and minimum utility for a group choice respectively (according to their own utility function).

It is also noteworthy that, in our analysis, users' utility functions are initially derived from their ratings, which are given independently from any group session, and then, they are updated to make them compliant with the preferences (in the form of like/dislike votes to the discussed items), which are disclosed during the group discussion.

Overall, the main contributions of this work are:

- We design a group discussion simulation model where the four considered factors can be manipulated as independent variables. In the simulation we have studied the effect of recommendations generated with three classical preference aggregation strategies, i.e., Average, Borda count and Multiplicative. The quality of the group recommendations is measured by observing the variations of two metrics (MIL and MMD) which are our dependent variables.

- Using the simulation model, we observe that:
  - When group members have similar preferences, groups of members with *accommodating* or *collaborating* styles get the lowest average utility loss, yet the difference in utility between the members is affected neither by the individual's conflict resolution style, group sizes nor the length of interactions.
  - When the groups are composed of users with diverse preferences but identical conflict-handling style, the average loss of *accommodating* and *collaborating* users decreases as more group interaction cycles are executed and it is much lower than in *competing* and *avoiding* groups. However, the utility discrepancy between winners and losers in the groups composed of *competing* and *avoiding* individuals is smaller than in groups formed by users with the other styles.
  - When a group contains *competing* users and ones with another style, it turns out that in the groups where there are also *accommodating* or *collaborating* users, the obtained average loss is the lowest, but the gap between the winner's and loser's utility is the widest and the winners are typically the competitive individuals.
- To further support the validity of the proposed simulation model, we analyse data collected from observing the decision making processes of groups of real users, which comprise of 27 participants organized in 8 groups faced with a travel decision task. The original findings of our simulation study can also be partially observed in real groups, i.e., the average loss of *collaborating* users is the lowest.

We believe that the knowledge of the effect of user conflict resolution styles on the quality of the group choice, which we have gained from this work, can inform and motivate follow-up on-line user studies. In fact, by automatically estimating the changes in users' preferences and classifying their conflict resolution styles a system can be designed to adapt the learned individual user models and its recommendation strategy to the estimated group situation, so that trade-offs between individuals and group benefits can be better balanced.

The rest of the paper is structured as follows. In Section 2, we give an overview of the related work. In Section 3, the interaction of a group discussion support system, the recommendation model and the observations of users' activity are illustrated. In Section 4, the proposed simulation model of the group discussion process is presented. The simulation experiments are elaborated in Section 5. In Section 6, we discuss the results obtained from the simulated group discussions, while in Section 7 we present the analysis of real groups, including data collection, methodology and its results. Finally, in Section 8 we provide conclusions of the study, and outline the future work.

2 Related work

In this section we summarize related research on preference aggregation for group recommendations; interactive GRSs; personality in groups; and evaluation methods of GRSs. We then position our contribution in relation to the state of the art.

2.1 Preference aggregation

To generate group recommendations, Jameson and Smyth [20] identified three alternative approaches: i) aggregation of the individuals' recommendations; ii) aggregation of the individuals' predicted ratings; and iii) aggregation of the individuals' profiles into a joint group profile.

To aggregate either profiles or predictions, Masthoff [27] gave a detailed discussion of preference aggregation strategies that are derived from *Social Choice Theory*. Overall, through empirical experiments, the results presented in the literature show that there is no best strategy since the effectiveness of a strategy depends on the characteristics of the application domain, the task and the group scenario [3, 7, 9]. This conclusion was also confirmed by an observational study on group decision processes [13, 15].

Different aggregation strategies have their own advantages and shortcomings since they are aimed at optimizing different criteria. The Average method, for instance, pursues fairness by equally considering all the group members' interests, and it is one of the most commonly used methods [20]. The Least-Misery strategy is useful for cases where some members have extreme preferences that can act as a veto (e.g., a vegetarian cannot eat meat). The Borda count method aggregates the ranked lists of alternatives given by each group member, whereas the Multiplicative strategy amplifies the differences between the group members' scores for the alternative options by multiplying them. The use of these strategies was further discussed by Masthoff [26].

2.2 Interactive group recommender systems

Another important line of research on GRSs has been dedicated to techniques for supporting recommendation processes that are more interactive. The user is supposed to be actively engaged in a sequence of user/computer interactions, such as eliciting their preferences, assessing the recommendations, and revising the elicited preferences in the light of newly acquired information.

One contribution in this direction was made by Jameson [19] with the *Travel Decision Forum*, a system that allows users to interact with embodied conversational agents representing group members, to agree on a set of shared preferences. *Trip@dvice*, developed by Bekkerman et al. [5], is a GRS that also exploits agents acting on behalf of group members and applies a cooperative negotiation methodology to tackle the group recommendation problem.

To date, critiquing is likely to be the most popular interactive approach. It enables users to interactively give feature-specific feedback on the recommended items and let them iterate the “recommend - review - revise” cycle until a desired item is found. The critiquing approach is well-illustrated by the *Collaborative Advisory Travel System – CATS*, designed by McCarthy et al. [29], a critique-based GRS that helps a group of users in planning a skiing vacation. Following this research direction, Guzzi et al. [18] introduced *Where2eat*, a mobile application for restaurant recommendations, which features “interactive multi-party critiquing”, an extension of the critiquing concept to computer-mediated conversations between two group members. Márquez and Ziegler [25] designed *Hootle+*, a hotel GRS that supports the discussion and negotiation of the features of the desired hotel. With a wider scope of applicability, Stettinger et al. [40] developed *Choicla*, a group decision support environment that allows users to provide feature-based feedback to flexibly configure decision tasks in a domain-independent setting.

In most of the mentioned work, users’ feedback revealed during the group decision making process is acquired in the form of feature-level critiques. Entering these critiques, however, tends to have quite a cognitive cost as users are expected to determine which feature values they like or dislike, which requires a great effort especially when the number of features is large [22, 30]. Moreover, so far, little attention in interactive GRSs has been paid to understanding how session-based preferences revealed during a group decision making process relate to the individual’s long-term preferences (i.e., the ones acquired outside of the group context). The aforementioned issues can be tackled by enabling the elicitation of session-based preferences expressed only at the item level. That is, to enable the user, in the context of a group discussion, to express only evaluations (like/dislike) for the proposed items, and leaving to the system the burden of understanding the implications of these preference statements on the relative importance of each single feature for the user (long-term preferences). This idea was originally introduced in [31] and is adopted by our recommendation model presented in the next Section 3.2.

Table 1 summarises the interactive GRSs that we have reviewed and classifies them according to two dimensions: (i) the group recommendation approach (profile or recommendation aggregation[1]), and (ii) the type of user’s preferences (long-term or session) used in each system. We reiterate that individual long-term preferences are independent from groups the user may join, whereas the session-based or group-induced preferences are related to the specific interaction with a group and a GRS in the course of a group discussion.

[1] The aggregation of rating predictions can also be referred to as the aggregation of recommendations [3, 27].

Table 1 Overview of the reviewed interactive GRSs

System	**Recommendation approach**		**Preferences**	
	Profile aggr.	Rec. aggr.	Long-term	Session
Travel Decision Forum [19]	✓		✓	
Trip@dvice [5]		✓	✓	
CATS [29]	✓		✓	✓
Where2eat [18]		✓		✓
Hootle+ [25]	✓			✓
Choicla [40]	✓		✓	

2.3 Personality in group recommendations and decision making

User personality is a fundamental and distinguishing feature of the user. It is studied in social psychology and is known to be strongly correlated with the group decision making process and its outcome [16, 42].

In the scope of this work, we focus on the conflicting interactions between users, hence we are interested in modelling personality dimensions that relate to this type of behaviour. In this context, the Thomas-Kilmann Conflict Model (TKI) has been formulated in an attempt to rationalize how a person behaves in conflict situations [21]. According to TKI, the behaviour of an individual in a conflict situation can be described by two dimensions: assertiveness and cooperativeness that are, respectively, the extent to which the individual attempts to satisfy her own interests or the interests of other people. These basic dimensions of behaviour are used to define five different conflict management styles that people can follow: *competing* (assertive and uncooperative), *accommodating* (unassertive and cooperative), *avoiding* (unassertive and uncooperative), *collaborating* (assertive and cooperative), and *compromising* (moderately assertive and cooperative). The conflict resolution style of a user is typically assessed by using questionnaires [21] or derived from the user's Big Five personality scores [46].

The role of conflict resolution styles, in the context of group recommendations, has been investigated only to some extent. Recio-Garcia et al. [36] took into account the group members' personality, modelled according to TKI, to weigh the influence of their preferences during the recommendation process; stronger influence was given to the more assertive members. This approach was extended by Quijano-Sanchez et al. [35], where the authors considered both personality strength and trust between group members to formulate an influence-based rating prediction. The predicted preferences of individuals in a group context hence depend, not only on their personality strength, but also on their trust for the other group members. Also inspired by the TKI model, Rossi et al. [39] proposed a negotiation mechanism for group recommendations, where the behaviour of agents representing group members is determined by the conflict resolution styles.

Regarding the group decision making process, Delic et al. [14] conducted an observational study whose results indicate that individual satisfaction with the group choice is strongly related to personality and conflict resolution styles.

Specifically, it was found that satisfaction with the choice was positively correlated with the *Agreeableness* and *Conscientiousness* personality traits and negatively correlated with *Neuroticism.* Additionally, it was observed that *collaborating* or *accommodating* groups are significantly more satisfied, even when their initial preferences mismatch the resulting group decision. It is worth noting that in the research study [14] the individuals' conflict resolution styles were derived from participants' personality scores measured by the Big Five model. Conversely, in our work, the conflict resolution styles are simulated. For each style, we have defined the simulated users behaviour by trying to match the definitions in the TKI with the actions that users have available to them in the chat-based interaction with the other group members and the recommender system. So for instance, a competing style is simulated as a user that often proposes items that he or she likes and tends to give negative evaluations to items proposed by the others.

With a user study, Barile et al. [4] additionally indicated that there is a direct correlation between tie strength and positive opinion shifting; i.e., in the presence of a peaceful relationship, the initial opinion of an individual shifts towards the opinion of another person. On the other hand, the presence of conflict leads to the opposite effect: individual's opinions drift further apart.

2.4 Evaluation of group recommendations

In order to evaluate the effectiveness of a group recommendation model both user studies and off-line experiments have been conducted [44]. Nevertheless, how to properly evaluate a GRS is still a major research topic [27, 12]. We discuss below two approaches.

2.4.1 User studies

A user study is carried out when the criteria used to measure the system performance are related to system usability and user experiences (e.g., perceived user's satisfaction or recommendation quality) [7, 15, 18, 40]. This type of study can be performed by directly interviewing participants or through crowd sourcing sites such as Amazon Mechanical Turk [4]. This approach, however, cannot be the sole method for evaluating the efficacy of a GRS as it is problematic to extensively test the system performance across many different group compositions, especially when dealing with a high degree of interactivity.

2.4.2 Off-line evaluations

As in classical RSs, off-line evaluations are also used in GRSs research. However, these approaches are somewhat hindered by the absence of public data sets that capture the preferences of users in real group contexts. To this end, synthetic groups that are sampled from standard data sets such as MovieLens are often generated [3, 11]. This solution, however, is based on the underlying assumption that the preferences of individuals are stable and independent from the group decision making process, which is actually not the case in most scenarios. In fact, the opinions or judgments of users in a group are likely to be influenced by the other group members. This effect has been observed and categorized as emotional contagion and conformity [28]. As a result, specific approaches come into play, such as those previously employed to simulate the behaviour of users when they interact with a RS [8, 24, 45]. Masthoff and Gatt [28] also conducted a simulation experiment to understand the impact of their proposed satisfaction functions on predicting users' satisfactions.

In the context of interactive GRSs, we have already proposed a group discussion simulation model where the impact of alternative combinations of long-term and session-based preferences on the recommendation performance in different group scenarios can be studied [33]. With that model, we simulated three group scenarios: when users' preferences in the group context match those that the users expressed while evaluating items for their individual consumption (*independence*); when they adapt their preferences to the group context and try to move them to get closer to the other group members' preferences (*conversion*); and when they even further differentiate their preferences while interacting in the group (*anti-conformity*). This model, however, only focuses on the case where all group members adopt the same response (preference revision) to the group setting (i.e., either *independence*, *conversion* or *anti-conformity*), and where they have the same level of participation in the group discussion, which might rarely occur in real life. Therefore, in this paper we aim at further expanding our simulation model with more diverse group compositions and more realistic scenarios.

In summary, we believe that an effective GRS must be able to adapt to the individuals' preference that are revealed in a particular group situation, so that it can suggest a more appropriate group choice. Moreover, since the group situation can be characterized by various group factors, in this study we focus on the following ones: (i) the specific user's behaviour in conflict situations, (ii) the similarity among users' preferences, (iii) the length of the group interaction, and (iv) the number of group members.

From the literature review, we notice that the evolution of group members' preferences during the decision making process, and their inter-dependencies with these four factors, have not been examined sufficiently, neither in user studies nor in off-line experiments. Hence, with the proposed group discussion and choice simulation procedure, we can experimentally manipulate different group compositions and analyze the responses of group members as a function of their preferences and their conflict resolution styles. We argue that the proposed simulated environment can be beneficial to practitioners in the field of GRSs to test various properties of a group discussion before they conduct empirical studies in a real-world setting.

3 Group discussion support system

Research studies into the functional theory of group decision making show that the full decision process taken by a group, which consists of various steps, determines the quality of the final decision output [16]. This finding is bolstered by an observational study on group decision processes highlighting that users' preferences are constructed during the process of making group decisions [13]. With that lesson learned, we have developed an interactive mobile GRS that is equipped with various decision support functions such as group discussion, group recommendations and choice suggestion. They are aimed at facilitating the various stages of the group decision making process [32].

In the next sub-sections, we first describe the interaction that the system supports, and then we explain the group recommendation model that is implemented to provide recommended items for group members. Finally, we analyze the data collected in some group decision making sessions that have been supported by the system, in order to motivate the design of the group discussion simulation model introduced in Section 4.

3.1 Interaction with the system

The developed GRS is an Android-based application featuring a chat-based interface, called STSGroup (South Tyrol Suggests for Groups). It is an extension of STS [10], a context-aware point of interests (POI) recommender originally developed for individuals. The flow of user-system interaction in STSGroup is illustrated in Figure 1.

As a concrete example, let us assume a user is looking for a POI in South Tyrol (Italy) to visit together with a group of friends. After signing up or logging in, she can invite her friends to join a group discussion by sending connection requests. The members of this group discussion can then perform as many interaction cycles as they desire, where a cycle is defined as a POI proposal made by one group member, followed by possible feedback given by the others. After some POIs are discussed, the group is supposed to select one of the proposed POIs to visit.

The system provides a set of user functions for creating a group discussion, proposing items to discuss, and giving feedback on the items proposed by other members in the form of: *best choice* (crown icon), *liking* (thumb up), and *disliking* (thumb down), as shown in Figure 2(a). We note that when a user proposes an item to the group discussion, the system automatically considers it to be the user's best choice unless (s)he explicitly gives a different evaluation. It is important to stress here that the user's *best choice* is supposed to be the option the user thinks would be best for him or her in the group session. However, users will differ in how much their evaluations consider the other group members' needs and wants. As a part of the interaction, any member can also decide not to evaluate any of the discussed items.

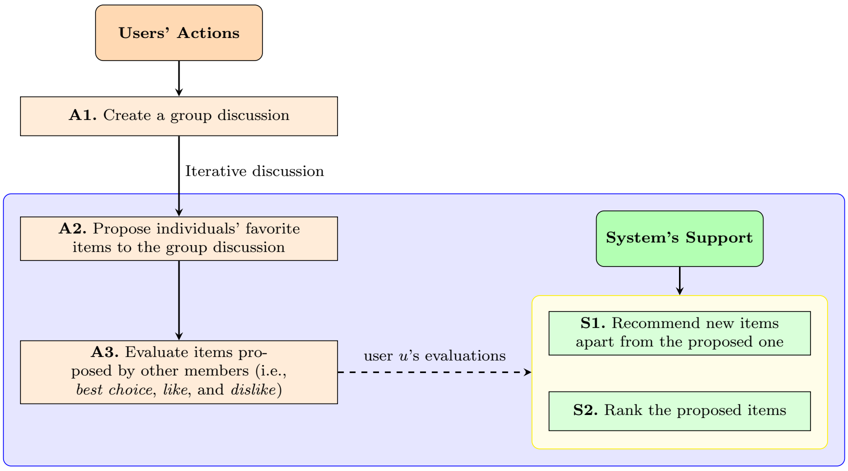

Users' Actions
A1. Create a group discussion
Iterative discussion
A2. Propose individuals' favorite items to the group discussion
A3. Evaluate items proposed by other members (i.e., best choice, like, and dislike)
user u's evaluations
System's Support
S1. Recommend new items apart from the proposed one
S2. Rank the proposed items

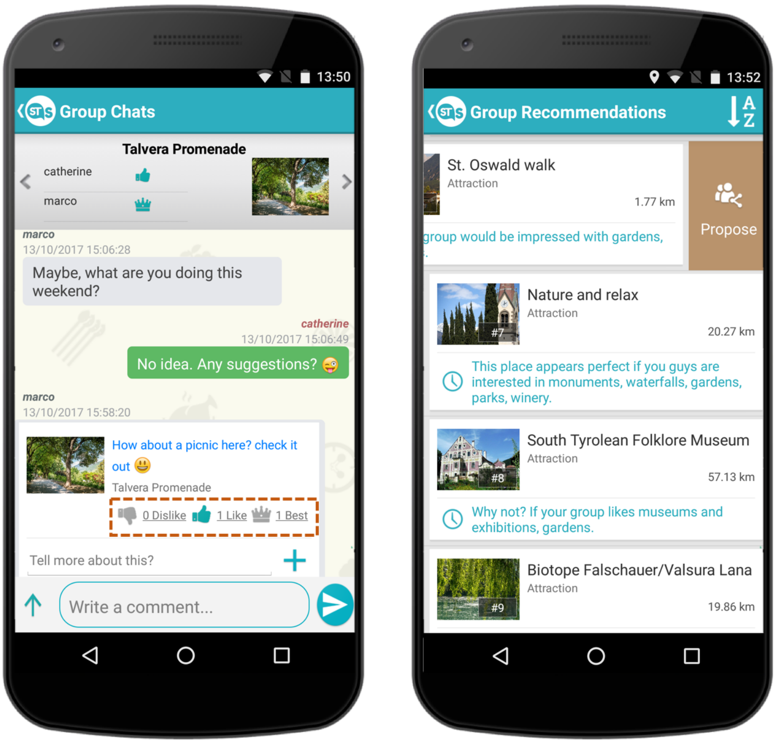
Fig. 1 A flow of users' interactions with the system

On the system side, by monitoring users' evaluations given in the course of the group discussion, the system continuously updates all the group members' preference models, which are implemented as utility functions, and have been previously initialized by considering the user group-unrelated ratings for POIs. The revised model is used to generate, upon the request of a user, novel group recommendations that enable group members to explore alternative options as new proposals (see Figure 2(b)). The system can also rank the discussed items and hence provide a final choice suggestion. Additional details about the recommendation model that is used to suggest new items from the entire item set or to rank the items proposed in the group discussion are presented in sub-section 3.2. After several interaction cycles, wherein POIs are proposed and rated by individuals, the discussion might end up with a group choice or without one.

3.2 Group recommendation model

Our group recommender system, before any group discussion unfolds, captures the individual user's preferences expressed in the form of ratings for POIs and builds a utility function for each user. Then the system continuously updates these utility functions by observing the evaluations revealed during the group discussion. More concretely, the recommendation model specifies how to:

1. represent items;
2. compute individual utility for items;
3. update the users' utility function during the discussion process; and
4. generate group recommendations.



13:50

Group Chats

Talvera Promenade

catherine

marco

marco
13/10/2017 15:06:28
Maybe, what are you doing this weekend?

catherine
13/10/2017 15:06:49
No idea. Any suggestions? 😜

marco
13/10/2017 15:58:20
How about a picnic here? check it out 😃
Talvera Promenade
0 Dislike 1 Like 1 Best

Tell more about this?

Write a comment...

13:52

Group Recommendations

St. Oswald walk
Attraction
1.77 km
group would be impressed with gardens,
.

Propose

Nature and relax
Attraction
#7
20.27 km
This place appears perfect if you guys are interested in monuments, waterfalls, gardens, parks, winery.

South Tyrolean Folklore Museum
Attraction
#8
57.13 km
Why not? If your group likes museums and exhibitions, gardens.

Biotope Falschauer/Valsura Lana
Attraction
#9
19.86 km

Fig. 2 Screenshots from left to right: (a) group discussion and (b) group recommendations

3.2.1 Item representation

Items are represented with Boolean vectors denoting the presence/absence of a set of keywords in their descriptions. This is a relatively simple approach, but it can give fairly good performance [23]. Specifically, item categories and full text descriptions are coming from a web-service provided by the Regional Association of South Tyrol's Tourism Organizations[2]. These are processed to extract keywords characterizing items. In the system, each item is modelled by the 10 most frequent features appearing in their categories and descriptions.

Hence, each item i is represented with an N-dimensional Boolean feature vector $x_i = (x_{i,1}, \ldots, x_{i,N})$, where N is the number of item features. If $x_{i,j} = 1$ then item i has the j-th feature, $x_{i,j} = 0$ otherwise.

3.2.2 Individual user's utility function

Before a group discussion starts, the system computes a utility vector w_u representing the importance that user u is estimated to assign to the N item

[2] http://www.lts.it

features. By using a content-based approach, the user's utility vector is built on the basis of the user's ratings acquired before the discussion, as follows:

$$w_{u,j} = \frac{\sum\limits_{i \in I_u} x_{i,j} r_{u,i}}{|\{i \in I_u : x_{i,j} \neq 0\}|}, \ j = 1, ..., N, \tag{1}$$

where $r_{u,i}$ is the rating that the user u gave for item i and I_u is the set of items rated by user u. By construction, if $w_{u,j} > w_{u,l}$ then the j-th feature is more important than the l-th feature according to user u. For example, let us assume that the ratings given by user u are ranging from 1 to 5 and are known for the following three items: $r_{u,12} = 3$, $r_{u,10} = 1$, and $r_{u,22} = 5$. Let us further suppose that the feature vectors of the considered items are: $x_{12} = (1, 1, 0)$, $x_{10} = (0, 1, 0)$, and $x_{22} = (1, 0, 1)$, respectively. Based on equation (1), the original (non-normalized) vector of feature weights w_u is computed as follows: $w_{u,1} = \frac{3+5}{2} = 4$, $w_{u,2} = \frac{3+1}{2} = 2$, $w_{u,3} = \frac{5}{1} = 5$. Then, w_u is normalized by dividing it by $\sum_{j=1}^{N} w_{u,j}$ (in the example's case 11), and $\sum_{j=1}^{N} w_{u,j}$ becomes equal to 1. Hence, the resulting (normalized) user utility vector is $w_u = (0.36, 0.18, 0.45)$.

Each user's preference model is then represented by a utility function:

$$U(u, i) = \sum_{j=1}^{N} w_{u,j} x_{i,j}. \tag{2}$$

The utility of an item gives a quantitative indication of the user's preference for the item (even those not yet rated), that is, we assume that a user prefers items with larger utility to those with smaller utility.

3.2.3 User's utility update

During the group discussion, we assume that the users will revise their utility functions taking into account (more or less) the full context of the group discussion. In practice, session-dependent constraints on the group members' utility functions are inferred from the revealed users' feedback (e.g., what they like and dislike) [43]. This means that, if a user favours item i over item i' then $U(u, i) > U(u, i')$ must hold[3]. It is worth noting that for each group member, the system can collect multiple constraints from their evaluations of items, so we denote with ϕ_u^g the set of constraints on the utility function of user u inferred from his or her evaluations given during the group g discussion. More generally, all items proposed in a group discussion can be classified (for each group member) into three sets: $BS(u)$ (*best choice*), $LS(u)$ (*liking*), and $DS(u)$ (*disliking*). As we stated previously, we assume that the user prefers items with larger utility, so the following constraints hold:

$$U(u, d) < U(u, l) < U(u, b) \tag{3}$$

[3] We make the simplifying assumptions that the features of items are independent of each other, and that group members are supposed to tell the truth about their preferences.

for all $d \in DS(u)$, $l \in LS(u)$, and $b \in BS(u)$. For example, let us assume $DS(u) = \{x_{25}\}$, $LS(u) = \{x_9\}$, and $BS(u) = \{x_8\}$. We assume also that the feature vectors of these items are $x_{25} = (1, 0, 0, 1, 0)$, $x_9 = (0, 1, 1, 0, 1)$, and $x_8 = (0, 1, 1, 1, 0)$, respectively. Then the constraints $U(u, 9) > U(u, 25)$ and $U(u, 8) > U(u, 9)$ are inferred. As a result, by considering the items' vector representations, the system derives that $\phi_u^g = \{c_1 : \{w_{u,2} + w_{u,3} + w_{u,5} > w_{u,1} + w_{u,4}\}; c_2 : \{w_{u,4} > w_{u,5}\}\}$.

It is important to note that we also check whether a newly-generated constraint c_u^g, on the user u utility in group g, is consistent with his/her existing constraints in ϕ_u^g. Precisely, we perform a partial check in order to speed up the computation, which consists of looking for a solution (utility vector) that satisfies the pairs of inequalities, c_u^g and c_j, for each $c_j \in \phi_u^g$. If there is no solution for one of these pairs of constraints we simply remove the old one c_j in the pair; in other words, our model prioritizes recent feedback. For instance, assuming that based on Equation 3 the model infers a new constraint $c_u^g : w_{u,4} < w_{u,5}$. One can see that the two inequalities c_2 and c_u^g have no common solutions. Hence, the model eliminates the constraint c_2 and adds the new one c_u^g to the set of constraints ϕ_u^g. However, it is still possible, even if it was never observed in our simulations, that a new constraint is compatible with each of the already acquired constraints but the full set of old constraints and the new one has no solutions. In this case we decided to not consider the new constraint.

By using the constraints in ϕ_u^g the system can update the previously computed users' utility functions in order to incorporate the users' preferences arising from the group interaction, hence aggregating long-term preferences with session (group specific) ones. We decided to update the utility vector of a user in such a way that it satisfies the constraints derived from the user's evaluations made during the group discussion and so that it is as close as possible to its original definition. The resulting optimization problem[4] is therefore formulated as follows:

$$w_u^g = \arg\max_{w} \cos(w, w_u) \text{ subject to } w \text{ satisfies } \phi_u^g, \tag{4}$$

where w_u^g is the updated utility vector of user u in the group discussion g.

3.2.4 Group recommendation

After having updated the group members' utility functions, the utility score of user u for each item i in the group discussion is re-calculated:

$$U^g(u, i) = \sum_{j=1}^{N} w_{u,j}^g x_{i,j}. \tag{5}$$

[4] In this work, the problem is solved by an off-the-shelf R package named *cccp*.

The group utility score[5] for an item, denoted by $U(g,i)$, can then be generated by applying several preference aggregation strategies. Specifically, in this paper we have used three of them, i.e., *Average*, *Borda count* and *Multiplicative* (this choice is justified by their fairness, ranking- or score-oriented characteristics as explained in Section 2.1).

Average: In this aggregation method, the updated utility functions of the group members are averaged.

$$U(g,i) = \frac{1}{|g|}\sum_{u\in g} U^g(u,i). \tag{6}$$

Borda count: Here items are independently ranked by using the updated utility functions of the group members $U^g(u,i)$. Hence each item will have a user dependent rank denoted as $rank_{u,i}$, where the top rank, 1, is taken by the item with the largest user utility. The group score for each item is calculated as the sum of the reciprocal of the individual rank:

$$U(g,i) = \frac{1}{|g|}\sum_{u\in g} 1/rank_{u,i}. \tag{7}$$

Multiplicative: The updated utility scores of group members are multiplied to generate the group utility for an item.

$$U(g,i) = \prod_{u\in g} U^g(u,i). \tag{8}$$

Finally, the system recommends the items with the highest group utility, $U(g,i)$, and hence the generated recommendations are the same for all group members. In the experimental evaluation described later in Section 5, in order to assess the quality of a recommendation, we compare, the highest utility that an individual user can get with her utility for the group choice. The highest utility that a user can get is identified by computing the utilities of all the recommendations and taking the maximum. This is the utility that the user could obtain if she had to make a choice only for her, without taking into account the group.

Algorithm 1 illustrates the execution steps of the recommendation model.

3.3 Data observations and analysis

We hereby describe how we collected logs of real user interactions during their group discussions with STSGroup. In an earlier implementation we conducted a user study that involved 15 participants (students and colleagues) formed in 6 groups including three groups of two and three groups of three users [32]. In a second extended study we involved 27 participants organized in 10 groups

[5] This must not be considered a “rating”, as there is no group rating that can be collected. It is just an aggregated score that, as in other GRSs, can be used to rank items in a proper way for the group.

ALGORITHM 1: Group recommendation process

```
▷ Before the group discussion
foreach $u \in g$ do
    Compute $w_u$, the individual utility vector of user $u$ as in Eq. (1)
end
▷ During the group discussion
while there is an action performed by $u$ do
    Infer user's $u$ constraints $\phi_u^g$ on the utility vector $w_u^g$
    Compute the updated user's utility vector $w_u^g$ by solving the optimization
      problem as in Eq. (4)
    Compute the updated utilities $U^g(u,i)$, $\forall i \in I$ as in Eq. (5)
    Rank items based on the group score $U(g,i)$, $\forall i \in I$ by using the aggregation
      strategies, Eqs. (6), (7) and (8)
end
```

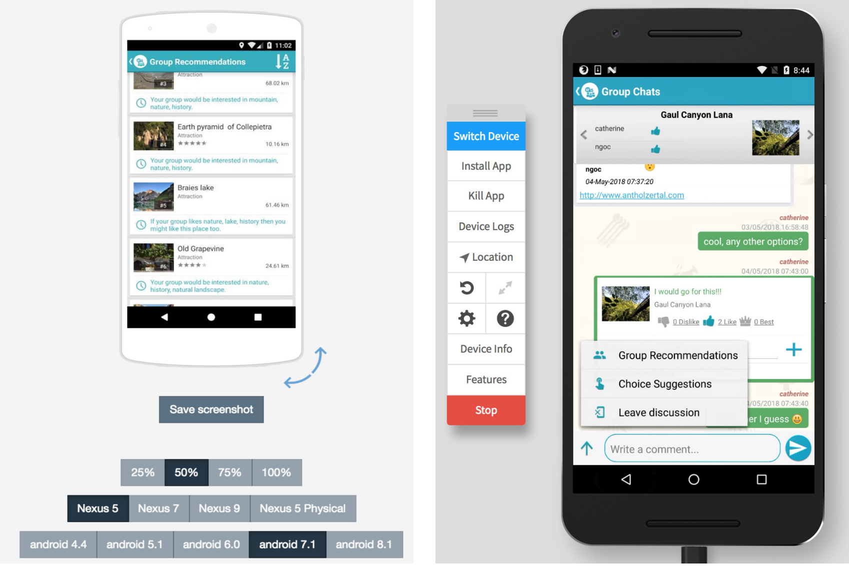
with four groups of two, five of three and one of four users. Both studies followed a three-phase structure: before, during and after a group discussion [32]. The only distinction, however, was that in addition to asking a set of subjects to meet and physically use STSGroup, the group discussion support system described in Section 3.1, the second data collection process was also conducted through the on-line mobile emulators Appetize.io[6] (Figure 3(a)) and BrowserStack[7] (Figure 3(b)) that allowed 4 groups consisting of three groups of two and one of three users to interact remotely with our web-based application.

In total 42 participants were engaged in 16 groups, i.e., seven groups of two, eight of three and one of four members. Before starting a group discussion, each participant was requested to freely choose at least five POIs to rate, so that their basic user profile could be acquired. More than 1000 POIs were considered, belonging to a wide range of sightseeing destination categories, such as natural monuments, historical buildings, castles, and so on. As we mentioned in Section 3.2.1, in the STSGroup, POIs are represented with Boolean vectors denoting the presence/absence of a set of keywords that are extracted from item descriptions coming from the web-service. The system, in the two studies together, collected 328 user's ratings for 120 POIs before any actual group discussion took place. Within the 16 group discussions mentioned above, 108 item evaluation were collected, i.e., the feedback was revealed in the form of either *best choice*, *like* or *dislike*. On average there were 2.7 interaction cycles for each group discussion. The longest group sessions terminated after 5 cycles while the quickest ones ended after 2 cycles. In our user studies, no specific request was made to the user about when to terminate the discussion. Hence, the discussion could end up whether or not the group choice was made.

When it comes to propose a POI (Action A2 in Figure 1), we observed in the user studies that the participants took turns to propose their preferred items, resulting in the number of items proposed by each user being nearly equal, i.e., they made approximately one proposal during each group discus-

[6] https://appetize.io

[7] https://www.browserstack.com/app-live/



Attraction
68.02 km
Your group would be interested in mountain, nature, history.

Earth pyramid of Collepietra
Attraction
★★★★★ 10.16 km
Your group would be interested in mountain, nature, history.

Braies lake
Attraction
61.46 km
If your group likes nature, lake, history then you might like this place too.

Old Grapevine
Attraction
★★★★ 24.61 km
Your group would be interested in nature, history, natural landscape.

Gaul Canyon Lana
catherine
ngoc

ngoc
04-May-2018 07:37:20
http://www.antholzertal.com

catherine
03/05/2018 16:58:48
cool, any other options?

catherine
04/05/2018 07:43:00
I would go for this!!!
Gaul Canyon Lana
0 Dislike 2 Like 0 Best

catherine
/05/2018 07:43:40
er I guess 😃

Fig. 3 Screenshots from the on-line mobile emulators that were used to collect data

sion. When it comes to assessing the items proposed by the other members (Action A3 in Figure 1), we noticed that users did not evaluate all the proposed items; in fact, the participants gave feedback only on 67% of the cases.

To understand the distribution of the users' utilities and how the utility of an item can influence the evaluation of it (best, liked, disliked or ignored), we calculated the possible utility values obtained by each user for the available POIs. Then, for each user, we took the 33rd and 66th percentile of the distribution of these utility values, so that the number of items was equally divided into three ranges of utility values. Then we computed the mean, among all the users, of the 33rd and 66th percentile utility values, which is respectively rounded up to 0.18 and 0.37, as visualized in Figure 4. Based on that analysis, we defined three ranges of utility values: *Low* $(0, 0.18]$, *Medium* $(0.18, 0.37]$ and *High* $(0.37, 1]$.

Interestingly, we observed that most of the time the participants proposed items with a large utility for themselves, hence assuming that these items would also be good for the group as a whole. This fact matches the utility maximization rule of utility-based choice theory [6]. This observation is important as well as supporting the definition and the usage of the user's utility function that we have defined.

On the contrary, but not surprisingly, given the results of the observational study presented in [13], we discovered that the participants did not always like the items proposed by other group members even though these items have the largest utility. Particularly, we estimated the probability that an item with a given user's utility value, that is, belonging to a particular utility range, would

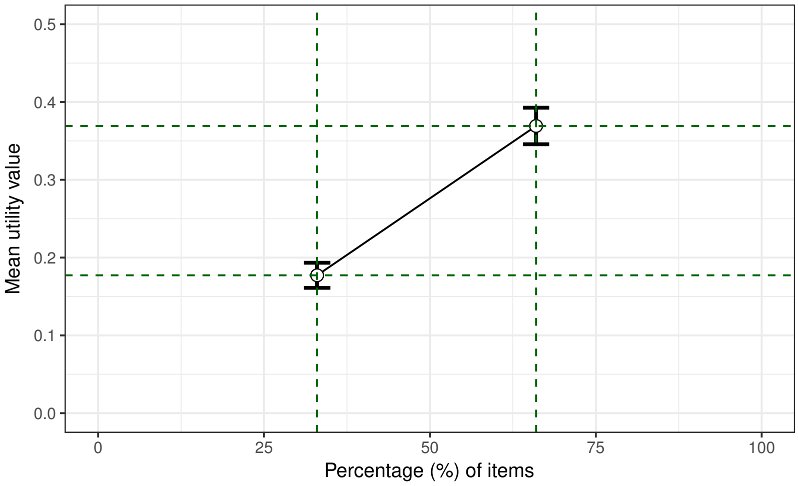


Fig. 4 The distribution of 33rd and 66th percentile user's utility values

receive a certain type of feedback (i.e., either *best choice*, *like*, or *dislike*). Based on the collected data, for each utility range, this probability is calculated as the proportion of the items that received that feedback among all types of observed feedback on items having utility in that range. Table 2 shows the partial randomness of individuals' feedback. For instance, let us assume that the utility value of user u for item 11 is 0.25, $U(u, 11) = 0.25$. This is an item of medium utility for the user. In this case the estimated probabilities that the item will receive the *best choice*, *like*, and *dislike* feedback from the user u are 0.28, 0.55 and 0.17, respectively. These observations are supported by the multinomial logit model, a probabilistic choice model, stating that the probability of choosing an alternative increases monotonically with an increase in the utility of that item and decreases with increases in the utility of each of the other items [34].

Table 2 Probability of each type of feedback in the three utility ranges: *Low* (0, 0.18], *Medium* (0.18, 0.37], and *High* (0.37, 1], which is determined by the ratio of the items that received that type of feedback (i.e., *best choice*, *like* or *dislike*) to all of the items observed in each range

Feedback	Probability		
	Low	Medium	High
Best choice	0.16	0.28	0.49
Like	0.31	0.55	0.38
Dislike	0.53	0.17	0.13

A further analysis of the logged information revealed that most of the group discussions (i.e., 11 out of 16) ended with at least one item that received most positive evaluations (e.g., *best choice* or *like*) from the group members, even though we did not explicitly ask the participants about their final group choice. We also observed that there were some discussions that finished without a clear choice being reached.

4 Group discussion simulation model

To find out how the four group factors, namely *conflict resolution style*, *inner-group similarity*, *length of group interaction* and *group size*, influence the quality of the group recommendation process, we have designed a group discussion simulation model which is justified and informed by the analysis of real users' interactions (see previous section). We will illustrate the simulation procedure by addressing the two main components of the procedure: (i) how group members behave and respond in conflict situations and (ii) how a group discussion proceeds. This is described in the following subsections.

4.1 Modelling individual behaviour in conflict situations

By reusing the TKI model [41], we identified different conflict resolution styles that group members can adopt. These styles are defined on the grounds of two fundamental dimensions: assertiveness and cooperativeness. In order to stick with the original TKI model we have implemented assertiveness as the intensity with which the simulated user expresses their own opinions, and cooperativeness as the extent to which the user tries to satisfy the other members' concerns. According to TKI, the four styles of dealing with conflicts are:

- *Accommodating* is unassertive and cooperative: individuals are inactive in pursuing their own concerns and try to satisfy those of the other members.
- *Competing* is assertive and uncooperative: individuals pursue their own concerns in an active way and refuse to accept proposals made by other members.
- *Avoiding* is unassertive and uncooperative: individuals are inactive in pursuing their own concerns, but also do not attempt to gratify those of the others.
- *Collaborating* is assertive and cooperative: individuals are active in following their own concerns and, at the same time, try to satisfy those of the other members.

In fact, the TKI model has a fifth style called *Compromising* that is moderate in assertiveness and cooperativeness but, for the sake of simplicity, we do not include this style, rather we examine a *Baseline* case, i.e., a simulated user who behaves like an average user in the analyzed STSGroup observations.

We decided to model assertiveness with the probability that group members propose items to the discussion with high utility, hence related to their concern. Hence, that probability is increased if a user is assertive and decreased

Table 3 Summary description of the adopted design for simulating the TKI conflict resolution styles

Style	Assertiveness	Cooperativeness	
	Propose	Evaluate positively	Evaluate negatively
Competing	$+$	$-$	$+$
Accommodating	$-$	$+$	$-$
Avoiding	$-$	$-$	$+$
Collaborating	$+$	$+$	$-$
Baseline	Use the probability derived from the STSGroup observations		

$+/-$ Increase / decrease the probability

otherwise. Conversely, cooperativeness is modelled by the probability of giving positive or negative evaluations to items proposed by other group members. Users with a cooperative style have a higher probability of giving positive feedback (i.e., *best choice* or *like*) and a lower probability of giving negative feedback (i.e., *dislike*). The opposite holds for the uncooperative users. A qualitative summary view of the model that we have implemented for simulating users' conflict resolution styles is provided in Table 3. On the basis of that, we implement the simulation procedure described later.

4.1.1 Assertiveness dimension

As we mentioned above, to simulate the intensity of user involvement in defending their own concerns, each user is characterized by a probability $p_p(u)$: the probability that in an iteration of a group discussion the user u will propose one of her favourite items to the group. The more assertive users are supposed to have larger $p_p(u)$ probabilities.

We considered the observed probability that a generic user of our user study made a proposal to their group as a baseline point. According to our data analysis (Section 3.3), the number of items proposed by participants was about equal. The probability that one group member makes a proposal in a group discussion iteration is therefore, in the baseline case[8], the same for all members $\forall u \in g, p_p(u) = \frac{1}{|g|}$.

Then, to simulate more or less assertive users in the simulation, that base probability is increased or decreased to some extent: $p_p(u) \pm \gamma$, $\gamma > 0$. In our experiments, the value of γ for each member in a group was a randomly chosen number in a range of [0.1, 0.2]. The probabilities are renormalized accordingly, by dividing them by $\sum_{u \in g} p_p(u)$, so that the sum of the probabilities of all group members is equal to 1.

4.1.2 Cooperativeness dimension

The cooperativeness level of each user is modelled in the simulation by increasing (with respect to the baseline probability) the probability that they

[8] There could be different ways to simulate turn-taking in proposing the items. In this work, we simply conjecture that users have the same probability $p_p(u)$.

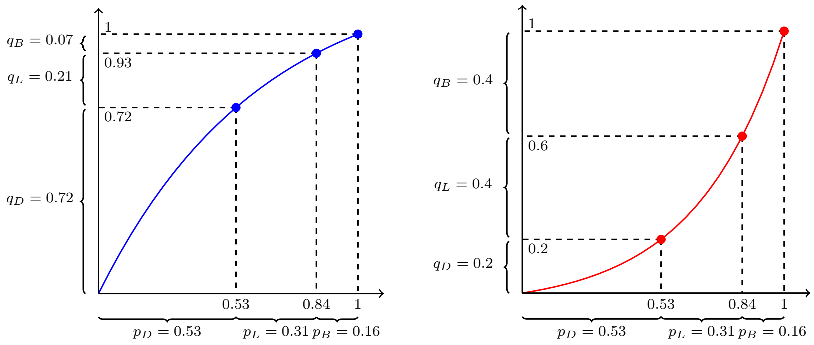
will give a positive evaluation (i.e., *best choice* or *like*) to the items proposed by other group members and diminishing the probability of giving negative feedback (i.e. *dislike*) to an item proposed by other users. For uncooperative users, the simulation applies the opposite strategy.

To transform the probabilities of giving a particular feedback given the estimated user utility for an item, we use an exponential cumulative distribution function (see Equation 9) that is illustrated in Figure 5(a) and 5(b).

$$f(x) = \frac{b^x - 1}{b - 1}, b \neq 1 \tag{9}$$

We set $x_1 = p_D$, $x_2 = p_D + p_L$, and $x_3 = p_D + p_L + p_B = 1$ where p_D, p_L and p_B are respectively the baseline probabilities that an item is assessed as *dislike*, *like* and *best choice*. The transformed values of p_D, p_L and p_B are denoted by q_D, q_L and q_B respectively ($q_D = f(x_1)$, $q_L = f(x_2) - f(x_1)$, and $q_B = 1 - f(x_2)$, so that $q_D + q_L + q_B = 1$). As we can see in the uncooperative case with $b < 1$, the function f increases q_D (i.e., from $p_D = 0.53$ to $q_D = 0.72$) and reduces q_L as well as q_B (i.e., from $p_L = 0.31$ to $q_L = 0.21$ and from $p_B = 0.16$ to $q_B = 0.07$). Conversely, in the cooperative case with $b > 1$, f decreases q_D (i.e., from $p_D = 0.53$ down to $q_D = 0.2$) and simultaneously increases q_L together with q_B (i.e., from $p_L = 0.31$ to $q_L = 0.4$ and from $p_B = 0.16$ to $q_B = 0.4$). We carried out the experiments described later with $b = 0.2$ and $b = 20$ for the uncooperative and cooperative case, respectively.

Since real users gave feedback on approximately 67% of the proposed items, each group member is simulated to evaluate an item proposed by the others with probability $p_f(u) = 0.67$.



(a) Uncooperative case (b=0.2)

(b) Cooperative case (b=20)

Fig. 5 An illustration of how the probabilities of giving *best choice*, *like*, and *dislike* feedback in the *Low* utility range can be transformed with respect to the level of cooperativeness

4.2 Modelling the group discussion

The overall structure of a simulated group discussion procedure is shown in Figure 6, which includes two logical components: *Group Simulator* and *Group Recommender*. Additionally, the blue box indicates steps associated with each discussion cycle (iteration).

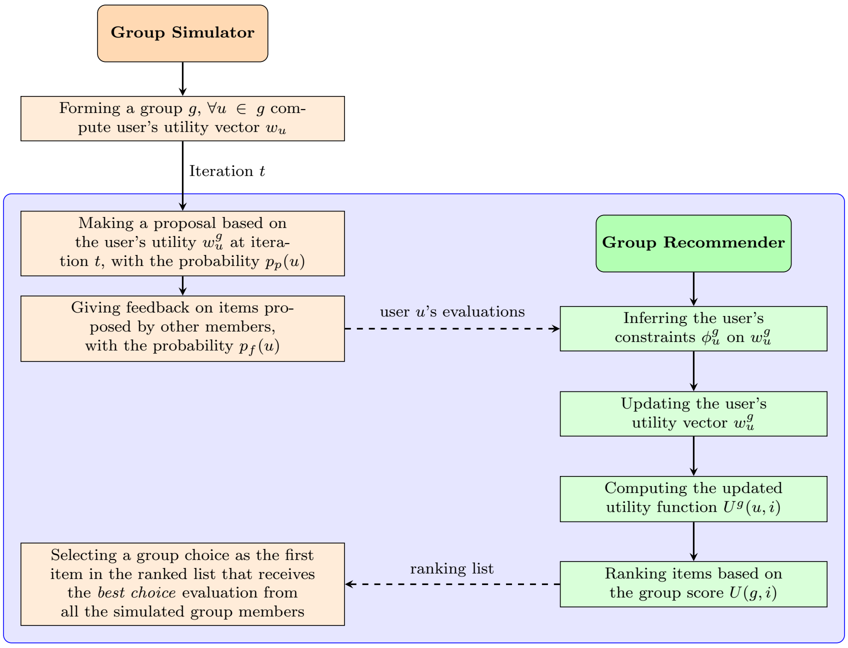

Group Simulator
Forming a group g, $\forall u \in g$ compute user's utility vector w_u
Iteration t
Making a proposal based on the user's utility w_u^g at iteration t, with the probability $p_p(u)$
Giving feedback on items proposed by other members, with the probability $p_f(u)$
user u's evaluations
Group Recommender
Inferring the user's constraints ϕ_u^g on w_u^g
Updating the user's utility vector w_u^g
Computing the updated utility function $U^g(u, i)$
Ranking items based on the group score $U(g, i)$
ranking list
Selecting a group choice as the first item in the ranked list that receives the *best choice* evaluation from all the simulated group members

Fig. 6 An illustration of the simulated group discussion process

The responsibility of the simulator boils down to the following sequence of steps:

1. ***Forming a group***. Due to the limited number of observed users, we have decided to generate synthetic users' profiles, which, however are similar to the profiles of the real users who evaluated STSGroup described in Section 3.1. More precisely, the simulator first estimates the real users' utility functions using Equation (1), based on their ratings elicited before any actual group discussions took place. Then, it applies k-means clustering[9] on the sample set of the real users' utility functions (utility weight vectors). The synthetic users' utility vectors, thereby, are drawn from a distribution with mean and variance of each cluster. This allows the simulation experiment to build groups composed of users with similar and different tastes (preferences). In the first case, the simulator arbitrarily chooses a cluster,

[9] In our experiments, we identified 5 clusters ($k = 5$) as this is the maximum number of users in a group that we considered.

and then generates the group members' utility vectors from this selected cluster definition (mean utility vector and variance). In contrast, in the second case, the simulator creates groups of users with diverse preferences by generating users' utility vectors from different clusters.

2. ***Making a proposal***. First, the simulator randomly selects a member and decides if this member makes a proposal with probability $p_p(u)$, whose value varies depending on the degree of assertiveness. If not, the selection is started again. Afterwards, the simulator identifies what the user will propose to the group discussion by considering the items with the highest user utility. It is noteworthy that the items proposed by each simulated member are selected according to their updated utility w_u^g, i.e., the best estimate of the session-dependent user utility.
3. ***Giving feedback***. The simulator then uses the probability $p_f(v), \forall v \in g$ and $v \neq u$, as a basis for selecting who among the remaining group members will give feedback on the item proposed by u. Each user's evaluation is determined by the cooperativeness of the specific conflict handling style modelled in Section 4.1. At each iteration t, the outcome of the simulator is a proposed item and the simulated users' evaluations for this item. This is used as input to the group recommender. The recommender then revises the users' utility models derived from the knowledge of their feedback revealed during the simulated group discussion at that iteration. It finally aggregates individual utilities or rankings to generate the group utility score $U(g, i)$, which is used to rank the discussed items and simulate a group choice.
4. ***Selecting a group choice***. At this point the simulator generates users' feedback (*best choice*, *like* or *dislike*) on the ranked items that were discussed, i.e., that were previously proposed by some users, and we assume that the collective choice is the first item in the ranked list that has received a *best choice* feedback from all of the group members. If none of the proposed items obtains this type of simulated feedback, then there is no group choice, and we assume that this produces no utility to the group members. Hence, in that case, the average individual loss – MIL, is taken as the mean utilities of the top individual choices (since the group choice has zero utility).
5. ***Holding a group discussion***. As part of our study is to understand how the number of iterations relates to the impact of conflict resolution styles on the group recommendation performance, the simulator considers two different settings: (i) the discussion runs for 10 iterations, and (ii) the discussion stops as soon as the first group choice is found. For the first setting, we analyze the group choice performance at various iterations (i.e., $t = 1 \dots 10$). We assume that the group choice at iteration t must be ranked at least equal to or higher than the previous choice at iteration $t - 1$, according to the group utility score $U(g, i)$. Otherwise a rational group should keep the preceding one. In other words, the group choice at a successive iteration must always improve a previous choice; no rational group should replace a previously made choice with another having inferior

utility[10]. For the second setting, the decision about how to simulate the stopping condition is informed by our observations of the discussion made in the user study wherein, most of the time, the group discussion ended, after some iterations, when one item received positive evaluations (e.g., *best choice* or *like*) from all the group members.

5 Experiments

We have run simulated group discussions and have generated data describing the user behaviour in these simulated situations. We analysed a number of alternative group settings defined by the independent variables: inner-group similarity, conflict resolution style, length of group interaction, group size and preference aggregation strategy. The dependent variable is the quality of group recommendations measured by two metrics: mean individual loss (MIL), and the difference between the maximum and minimum user's utility (MMD).

5.1 Independent variables

To obtain a clear picture of the impact of each analysed factor, we employed a factorial design, as summarized in Table 4, which shows the possible combinations of the values of the independent variables.

We considered two levels of inner-group similarity, namely homogeneous and heterogeneous groups where individuals have similar and diverse tastes respectively. As explained in Section 4.2 when describing how the groups were composed, groups of like-minded users are sampled from the same cluster while groups of members with varied interests are obtained from different clusters.

We studied the group discussion outcome in groups characterized by having a uniform conflict resolution style, i.e., where all group members have the same conflict resolution style, and also mixed groups, where users with a competitive attitude are mixed with users with another conflict resolution style. The five uniform groups are composed of users with conflict resolution styles: *competing*, *accommodating*, *avoiding*, *collaborating* and *baseline*. For the mixed-style groups, potentially there are many possible combinations, but as pointed out by studies in social psychology [16], competitive situations are likely to intensify conflict between individuals. We thereby focused on the situations involving competition. We specifically considered four mixed combinations in which *competing* simulated users are respectively paired with *accommodating*, with *avoiding*, with *collaborating*, and with *baseline* simulated users.

Regarding the length of group interaction, as we discussed earlier, we examined two scenarios: when the group discussion has an increasing number of iterations, up to 10; and when the group discussion is terminated as soon as

[10] This is based on the widely-used rationality assumption, i.e., that people make rational decisions to some extent [38]. Arguably, it is not always the case in practice.

Table 4 Overview of the employed independent variables

Independent variables	Number of levels	Values
Inner-group similarity	2	Homogeneous and heterogeneous groups
Conflict resolution style	9	5 uniform (*compete*, *accommodate*, *avoid*, *collaborate*, and *baseline*); and 4 mixed styles (*compete* & *accommodate*, *compete* & *avoid*, *compete* & *collaborate*, and *compete* & *baseline*)
Interaction length	2	Varying interaction length and stopping the discussion at the first group choice
Group size	4	Groups of size 2, 3, 4 and 5
Preference aggregation strategy	3	Average, Borda count and Multiplicative

a group choice is determined, i.e., when a discussed item is getting *best choice* feedback from all group members.

With respect to the group sizes and preference aggregation strategies, we studied groups of different size from 2 to 5 users and implemented three preference aggregation strategies Average, Borda count and Multiplicative.

Each of the considered group situations was evaluated in a series of 200 trials. On each trial, the discussion process described in Section 4.2 was performed. We report the average results of these 200 trials.

5.2 Dependent variables

The main goal of our research is to assess and understand the potential effect of the dynamic changes in users' responses prompted by different conflict resolution styles on the group recommendation performance in various group situations. We, therefore, inspect the quality of a decision making outcome supported by the recommendation process by measuring: 1) the variation in individuals' utility loss (i.e., *Mean Individual Loss* — MIL), and 2) the difference between the utility of a winner and a loser (i.e., *Max-Min Difference* — MMD), where winners and losers are defined as the simulated users that receive the maximum and minimum utility for the group choice respectively. The first metric is motivated by the idea of process losses in social psychology which measures the reductions in performance efficacy caused by group settings that hinder individuals from reaching their full potential [16]. This makes sense in the context of GRSs, as group recommendations that aim to simultaneously suit the preferences of all group members are hardly as good as those specifically tailored to individuals. Moreover, the occurrence of conflict between users' interests during the decision making process often ends up with one side branded the winner and the other side the loser, so this motivates our decision to observe the difference in utility between the winners and the losers that evolves during the simulated discussion.

5.2.1 Mean individual loss (MIL)

Mean Individual Loss (MIL) measures how much an individual loses in utility for the group choice with respect to their personal (optimal) choice. In particular, for each simulated user we estimate their individual loss by comparing their utility for the group choice i_g to their utility for the optimal individual choice i_u (the utility that a user can obtain by choosing the item with the highest utility). We take the average of all the group members' utility losses defined as in Equation 10:

$$MIL(g, i_g) = \frac{1}{|g|} \sum_{u \in g} U(u, i_u) - U(u, i_g). \tag{10}$$

It is important to recall that the personal choice is the item having the highest utility with respect to the original user utility vector w_u, the one that is built directly from the individual preferences independent of group settings. By studying the behaviour of this metric, we can judge the deviation from the personal optimal choice caused by the changes in users' behaviour that are also caused by their conflict resolution style.

5.2.2 Max-min difference (MMD)

This metric assesses the discrepancy between the winner's and the loser's utility for the collective choice. It is inspired by the definition of "satisfied winner" and "dissatisfied loser" coming from the prior research on individual satisfaction with group decisions in which the experience of winning (satisfied) or losing (dissatisfied) is defined by the match or mismatch between a group choice and the users' initial preferences [14]. Here in our case, the winner and the loser of a group decision making process are defined as those whose utility for the group choice is the highest and lowest, accordingly.

$$MMD(g, i_g) = \max_{u \in g} U(u, i_g) - \min_{u \in g} U(u, i_g). \tag{11}$$

6 Results

We hereby report the dynamics of the group recommendation performance measured by the metrics MIL and MMD for the group situations that we considered. For the sake of brevity, users that are, and groups composed by, *accommodating* or *collaborating* individuals are referred to as cooperative users and groups, respectively; while users that are, and groups composed by, either *avoiding* or *competing* individuals are called uncooperative users and groups, respectively.

6.1 Experiments with homogeneous groups

6.1.1 Uniform conflict resolution styles

In this section, we examine groups whose members have similar preferences and the same conflict resolution style.

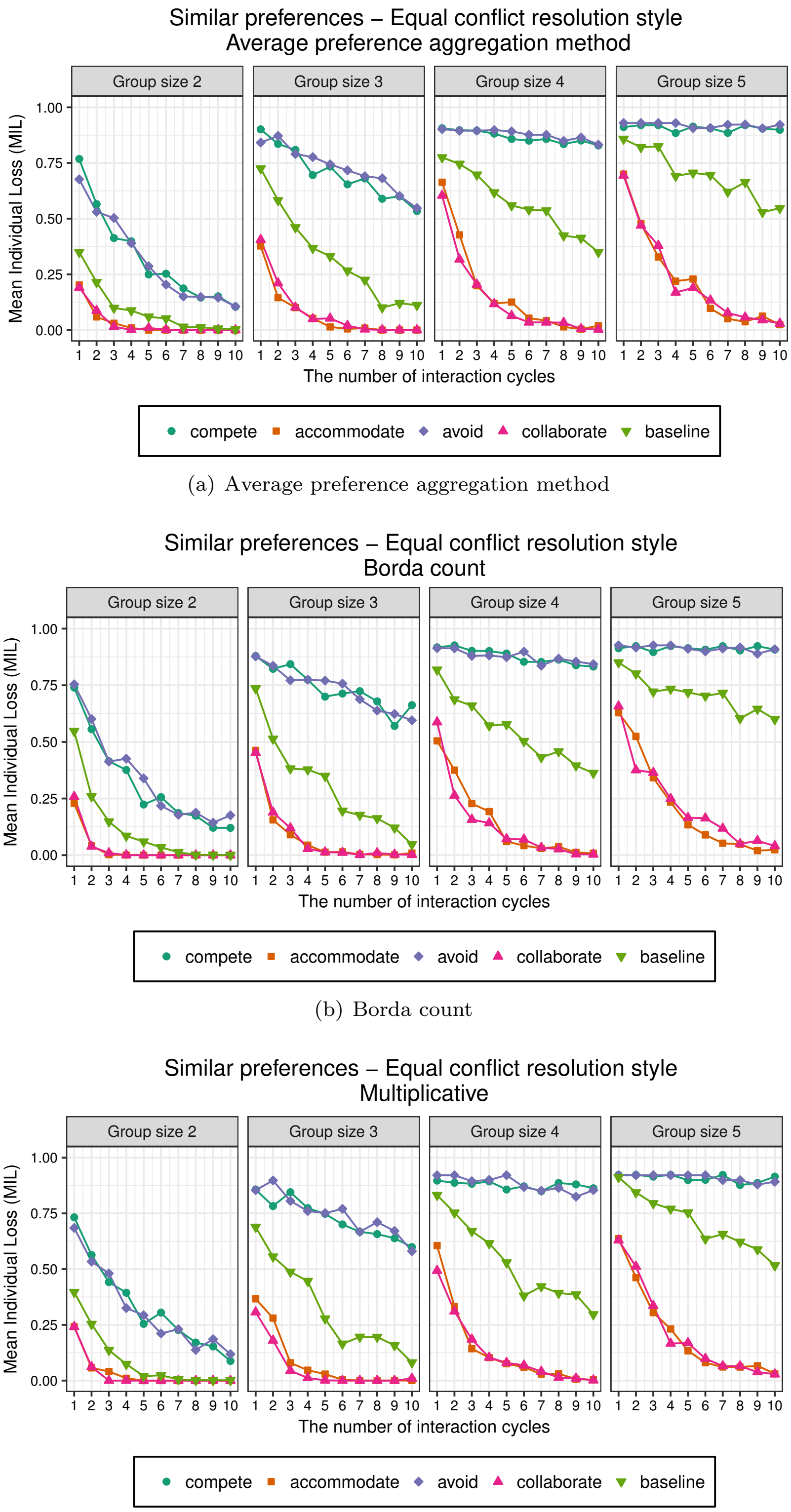
Varying the interaction length. Figure 7 displays the MIL for each conflict resolution style along $t = 10$ iteration cycles for the three preference aggregation strategies (Average, Borda count and Multiplicative) that we have considered. We observe that the average loss in cooperative groups is much lower than in uncooperative ones. In small cooperative groups (i.e., groups of size 2 and 3), group members are also able to find an ideal group choice, the item with MIL = 0, after few iterations, whereas this does not hold for uncooperative groups. Moreover, we see that large uncooperative groups have high MIL and the further iterations do not help in reducing it.

These results can be explained by the fact that *accommodating* and *collaborating* users are very likely to give the best evaluation to the items proposed by the other members, plus their tastes are similar in this case. Quite the opposite, there is a high chance that *avoiding* or *competing* members do not like the items suggested by the other members even when these suggestions are also fitting their preferences (i.e., they have rather high utility for them). Hence it turns out, they get the greatest utility loss, and their MIL hardly goes to zero at successive iterations. This is true especially in large group sizes (i.e., groups of size 4 and 5).

The simulation results obtained by using Borda count, Multiplicative and Average preference aggregation strategies are qualitatively similar. We thereby only report the obtained results for the Average preference aggregation rule from now on.

Figure 8 illustrates the MMD evolution; all these curves are constant and overlapping. This means that there is no difference in the utility of the winner and the loser, regardless of the user conflict resolution style, group size and number of iterations. This is quite clear because, as long as the group members have the same preferences they must not have different utilities for the same group choice.

To test whether or not the differences in MIL are statistically significant, we used a non-parametric Kruskal-Wallis test after the normality and homogeneity of variance assumption were not satisfied, respectively checked with Shapiro-Wilk and Levene's test, to perform a one-way ANOVA test. The results show the significant differences between the conflict resolution styles in each group size ($p < 2.2e^{-16}$). From the output of the Kruskal-Wallis test, we continued to calculate Wilcoxon rank sum test with Bonferroni correction for pairwise comparisons between the five conflict resolution styles. The tests performed in all group sizes indicate that most of the comparisons are significant ($p < 2.2e^{-16}$) except for the two comparisons between *accommodating* and *collaborating* ($p = 1$) and between *competing* and *avoiding* ($p = 0.78$). With



(a) Average preference aggregation method

(b) Borda count

(c) Multiplicative

Fig. 7 MIL of groups with similar interests and uniform conflict resolution styles

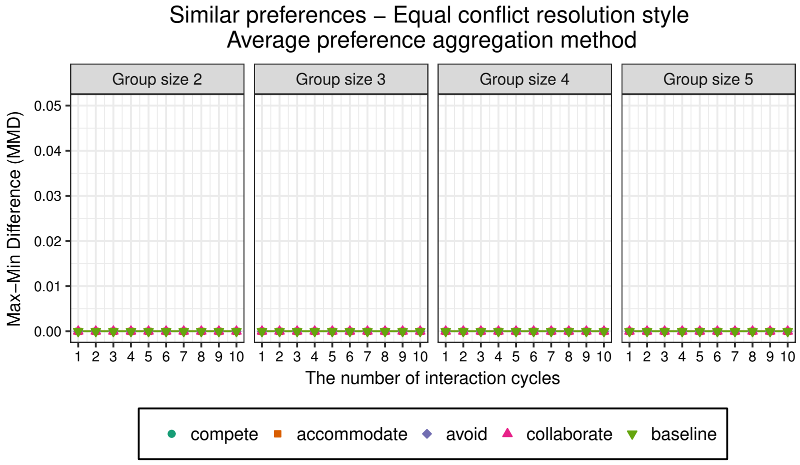


Fig. 8 MMD of groups with similar interests and uniform conflict resolution styles

respect to MMD, the statistical tests also confirm no significant difference between conflict resolution styles.

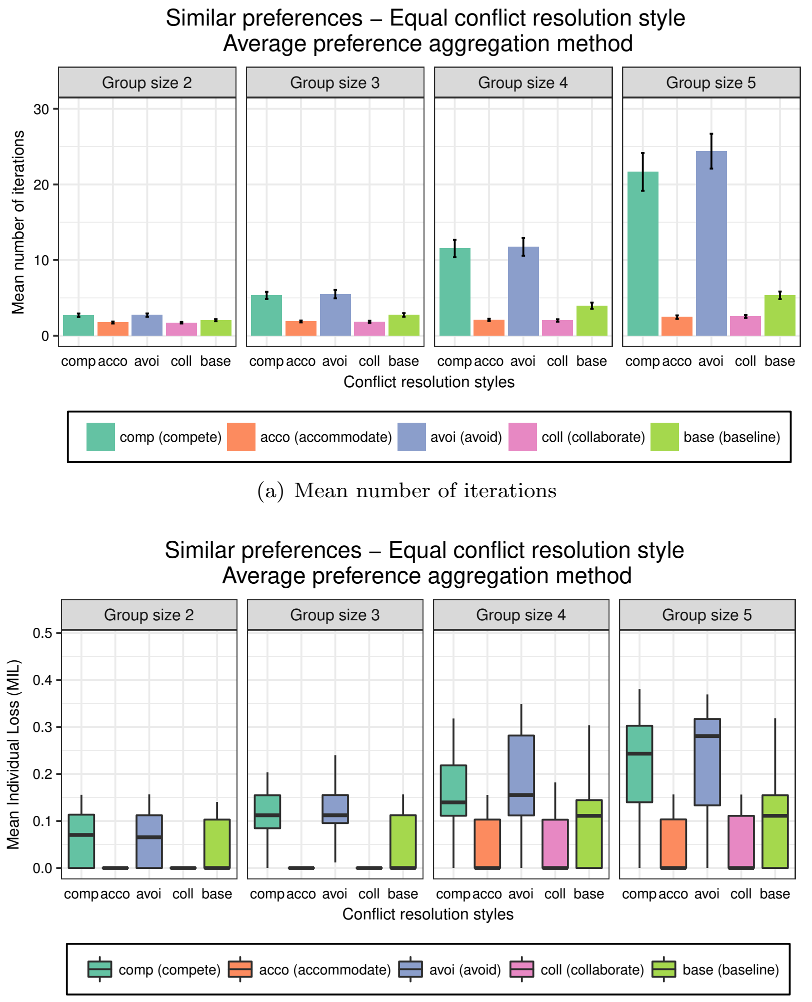
Stopping the discussion at the first group choice. In this setting, we analyze the quality of the first acceptable group choice. This means that the simulated group discussion stops as soon as the first choice is found.

Compared to cooperative groups, uncooperative groups, as expected, require more iterations to find a collective choice and their average loss in terms of MIL is also higher. These results are illustrated in Figure 9(a) and 9(b), accordingly. We also notice that the larger the group size is, the more iterations are needed to find the first group choice.

The Kruskal-Wallis test, again, confirmed that the difference in MIL between the considered conflict resolution styles is significant ($p < 2.2e^{-16}$). The results obtained from the pairwise comparison using the Wilcoxon test also lead us to the same conclusions as the previous setting (see Table 5 for the results of groups of size 2).

Table 5 The Bonferroni corrected p-values for pairwise comparisons in terms of MIL in groups of size 2 whose members have similar interests and uniform conflict resolution styles, stopping the simulation at the first group choice

	compete	accommodate	avoid	collaborate
accommodate	$1.4e^{-8}$	-	-	-
avoid	1.00	$5.5e^{-8}$	-	-
collaborate	$2.2e^{-10}$	1.00	$8.4e^{-10}$	-
baseline	0.0031	0.0417	0.0126	0.003



(a) Mean number of iterations

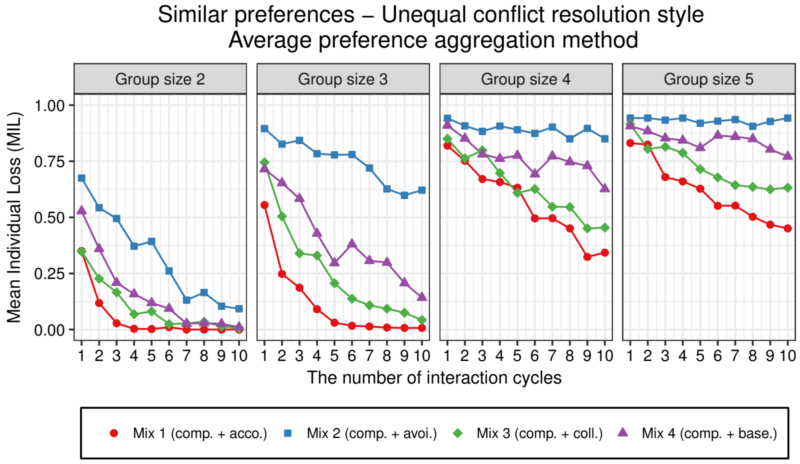
(b) MIL

Fig. 9 Performance of groups with similar interests and uniform conflict resolution styles, stopping the discussion at the first group choice

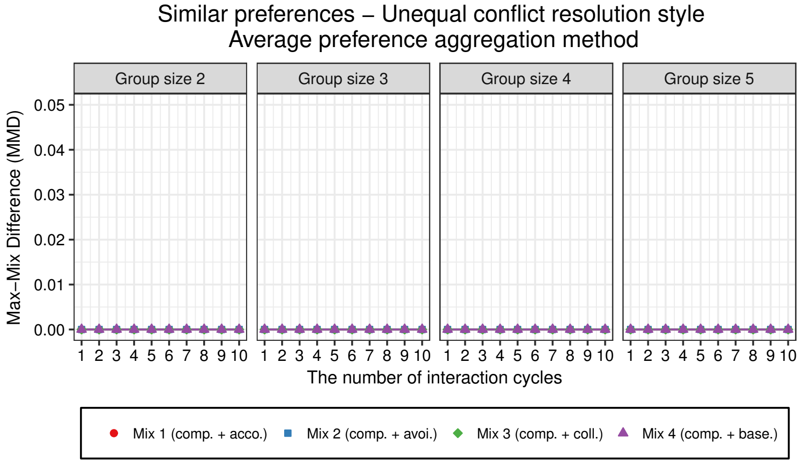
6.1.2 Mixed conflict resolution styles

We now examine the case when groups are composed of users with similar preferences, yet different conflict resolution styles.

Varying the interaction length. We find that mixed groups of *competing* and *accommodating* members have the lowest loss while the highest loss is scored by mixed groups of *competing* - *avoiding* users (see Figure 10(a)).



(a) MIL

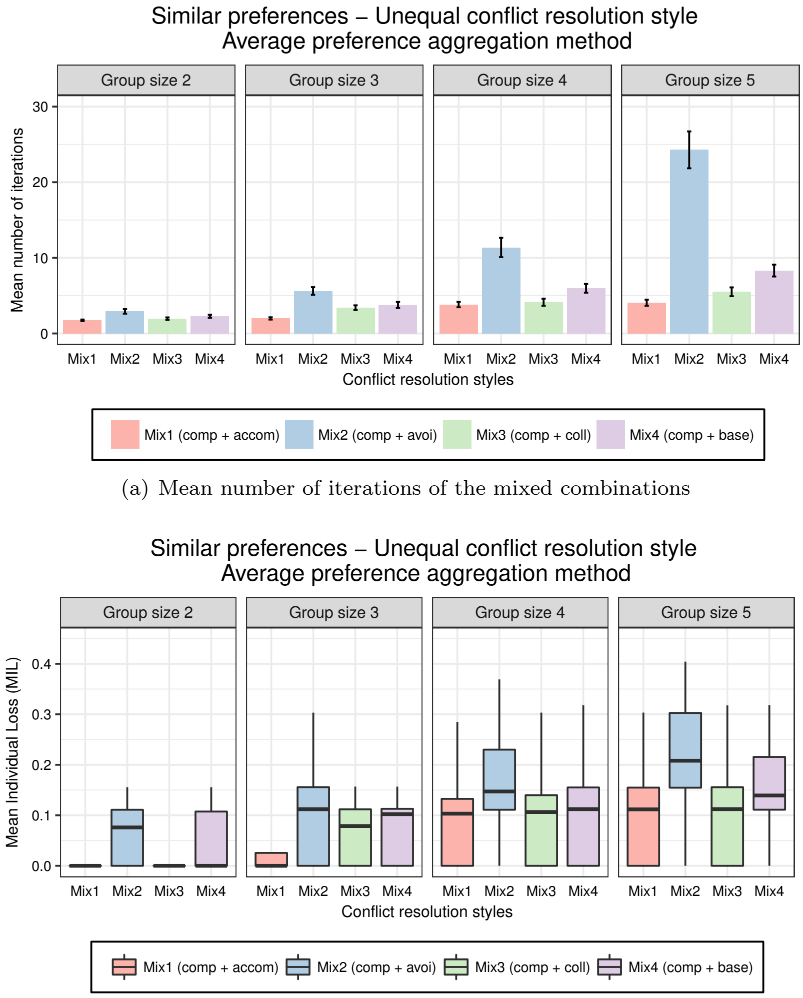


(b) MMD

Fig. 10 MIL and MMD of groups with similar interests and mixed conflict resolution styles

Regarding MMD, Figure 10(b) shows that the difference between the maximum and minimum utility of the members in a group is constant despite the varying combination of conflict resolution styles, group sizes and number of iterations. This result has the same explanation that we formulated for the previous setting.

The Kruskal-Wallis test results show that the difference in MIL of group members from mixtures of conflict resolution styles are significant ($p < 2.2e^{-16}$). The Wilcoxon pairwise comparisons corroborate the significant effect of mixed groups in all the group sizes ($p < 2.2e^{-16}$). Similarly to the uniform case, the statistical tests also point out that there is no significant difference between conflict resolution styles in terms of MMD.



(a) Mean number of iterations of the mixed combinations

(b) MIL of the mixed combinations

Fig. 11 Performance of groups with similar interests and mixed conflict resolution styles, stopping the discussion at the first group choice

Stopping the discussion at the first group choice. Figure 11(a) shows that mixed groups of *competing* and *avoiding* need more iterations, compared to the other combinations, in order to find the first group choice. Unsurprisingly, they also get the largest average loss, particularly in large group sizes (see Figure 11(b)).

Table 6 The Bonferroni corrected p-values for pairwise comparisons in terms of MIL in groups of size 2 whose members have similar interests and mixed conflict resolution styles, stopping at the first group choice

	Mix 1	Mix 2	Mix 3
Mix 2	$2e^{-7}$	-	-
Mix 3	0.53	0.001	-
Mix 4	$1.9e^{-5}$	1.00	0.028

Table 7 Summary results of the experiment with the homogeneous groups

Interaction length	TKI styles	MIL	MMD	# Iterations
Varying length	Cooperative groups (*collaborate* or *accommodate*)	smallest	0	1..10
	Uncooperative groups (*compete* or *avoid*)	largest	0	1..10
	Mix (*compete* & *accommodate*)	smallest	0	1..10
	Mix (*compete* & *avoid*)	largest	0	1..10
Stopping at the first group choice	Cooperative groups	smallest	largest	smallest
	Uncooperative groups	largest	smallest	largest
	Mix (*compete* & *accommodate*)	smallest	largest	smallest
	Mix (*compete* & *avoid*)	largest	smallest	largest

The larger the group size, the greater the effect.

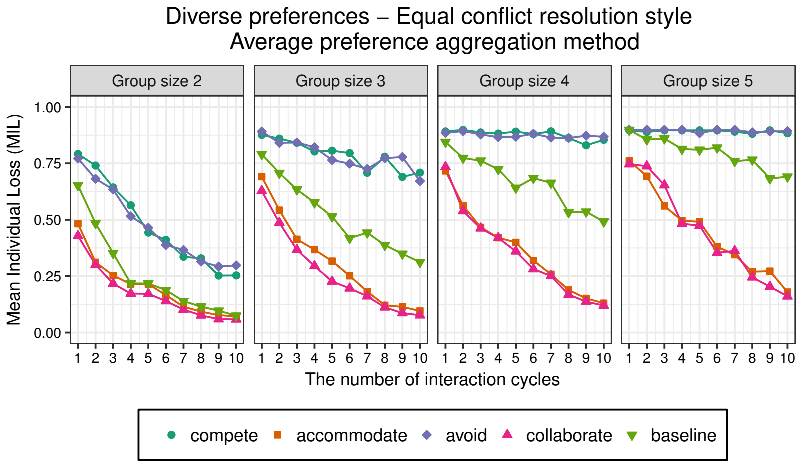
In this setting, the significant differences in terms of MIL are found with $p = 2.4e^{-8}$, $p = 7.6e^{-11}$, $p = 1.6e^{-8}$, and $p = 1.2e^{-12}$ for groups of size 2, 3, 4 and 5, respectively. Comparing MIL of mixed groups of size 2 in pairs, we observe that there are no significant differences between Mix 1 (*competing* and *accommodating*) and Mix 3 (*competing* and *collaborating*) as well as between Mix 2 (*competing* and *avoiding*) and Mix 4 (*competing* and *baseline*) (see Table 6). Similarly, the difference between Mix 1 and Mix 3 is not significant in groups of size 4 and 5, but the difference is significant for groups of size 3 ($p = .000$). We also find out that there is no evidence that the MIL for Mix 3 (*competing* and *collaborating*) is statistically different from the MIL of Mix 4 (*competing* and *baseline*) in groups of size 3, 4 and 5.

Table 7 summarizes all the results obtained from the homogeneous case wherein group members have similar preferences.

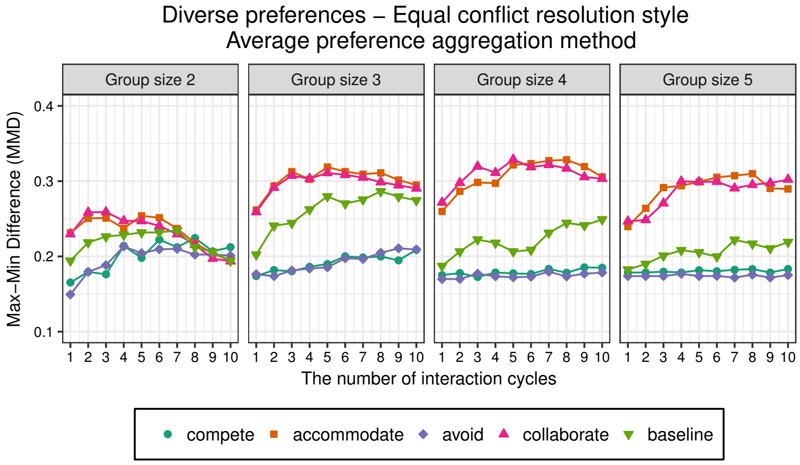
6.2 Experiments with heterogeneous groups

6.2.1 Uniform conflict resolution styles

In this section, we report the simulation results when the groups are composed of users with diverse preferences yet all the group members have the same conflict resolution style.



(a) MIL



(b) MMD

Fig. 12 MIL and MMD of groups with diverse interests but similar conflict resolution styles

Varying the interaction length. Similarly to the experiment with groups having similar preferences, we also observe in this case (Figure 12(a)) that the average loss, in terms of MIL, of cooperative groups is the lowest, while the highest loss is obtained by uncooperative groups. The reason is because members who work together cooperatively are inclined to accept the proposals of each other in a reciprocal way even though they might differ from their personal choice. Conversely, uncooperative users who are mostly focusing on their own concerns and reject alternatives proposed by other members, often end up with a group choice that is neither pleasing the others nor themselves.

Unlike the previous case wherein group members have similar preferences and the MIL of cooperative users tends to zero very quickly, in this case, the convergence to zero is a bit slower, and evidently, in the 10 observed iterations it never goes to zero. As expected, it takes more time to build acceptable recommendations for groups composed of individuals having diverse preferences, and reaching a consensus requires even more time, with uncooperative members or with larger group size.

Interestingly, Figure 12(b) shows that the MMD of uncooperative groups, i.e., the difference between the winner (maximum utility) and the loser (minimum utility), is the lowest. The uncooperative members who refuse to accept items proposed by the other group members lose more than cooperative users, but they all lose the same as each other (small MMD). Conversely, cooperative groups, which are composed of users that tend to accept proposals made by other users, even if they do not fit perfectly their own preferences, tend to collectively lose less utility but there are larger differences in utility between winners and losers (larger MMD).

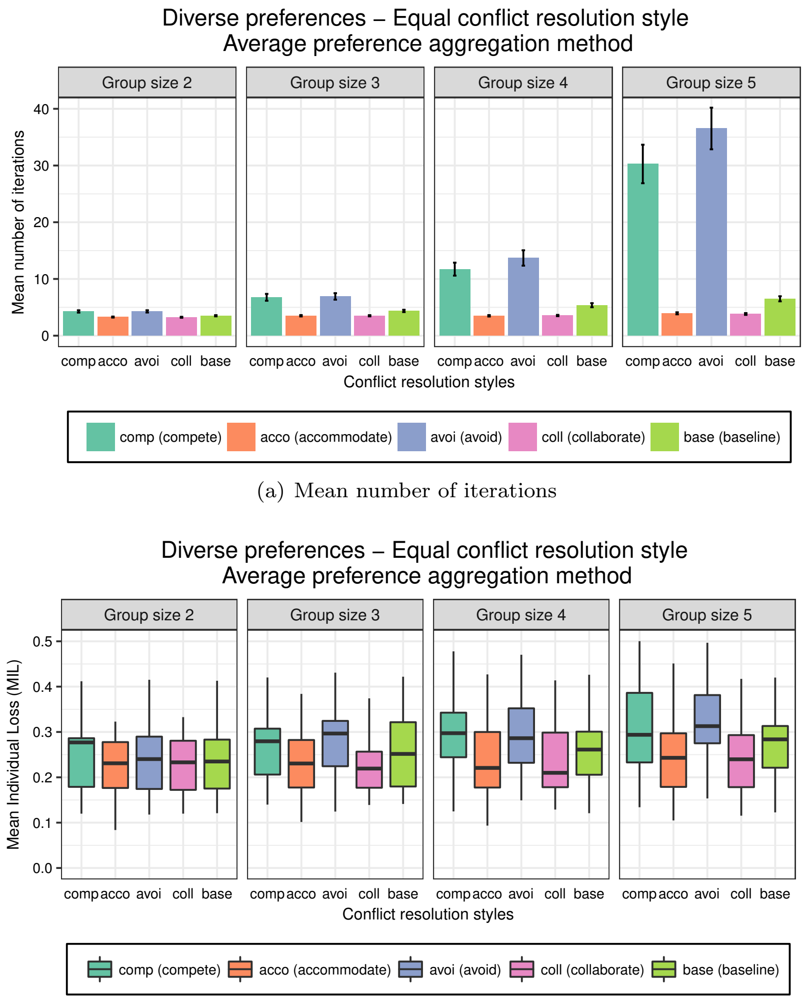
The Kruskal-Wallis test confirms that the results obtained from different conflict resolution styles in terms of MIL and MMD are significantly different ($p < 2.2e^{-16}$). With respect to MIL, it can be seen from the results of groups of size 2 shown in Table 8 that the conflict resolution styles have significantly different performances, with Bonferroni adjustment $p < 0.05$ (similarly to the other group sizes). Regarding MMD, however, we do not find a significant difference between *competing* and *avoiding* as well as between *accommodating* and *collaborating* in each group size (see Table 9 for groups of size 2).

Table 8 The Bonferroni corrected p-values for pairwise comparisons in terms of MIL in groups of size 2 whose members have diverse interests but similar conflict resolution styles

	compete	accommodate	avoid	collaborate
accommodate	$<2e^{-16}$	-	-	-
avoid	0.004	$<2e^{-16}$	-	-
collaborate	$<2e^{-16}$	0.013	$<2e^{-16}$	-
baseline	$<2e^{-16}$	$<2e^{-16}$	$<2e^{-16}$	$<2e^{-16}$

Table 9 The Bonferroni corrected p-values for pairwise comparisons in terms of MMD in groups of size 2 whose members have diverse interests but similar conflict resolution styles

	compete	accommodate	avoid	collaborate
accommodate	$<2e^{-16}$	-	-	-
avoid	0.31	$<2e^{-16}$	-	-
collaborate	$<2e^{-16}$	0.56	$<2e^{-16}$	-
baseline	$<2e^{-16}$	$<2e^{-16}$	$<2e^{-16}$	$<2e^{-16}$



(a) Mean number of iterations

(b) MIL

Fig. 13 Performance of groups with diverse interests but similar conflict resolution styles, stopping the discussion at the first group choice

Stopping the discussion at the first group choice. Figure 13(a) shows the average number of iterations needed to identify the first group choice. Similarly to the previous case, where the group members have similar preferences, uncooperative groups require more iterations for arriving at the group agreement, compared to cooperative groups. It is also observed that groups whose members have diverse preferences need more iterations to find this first group choice, compared to groups composed of members having similar preferences. Moreover, groups composed of uncooperative users are unable to find a group choice before a very large number of iterations. In fact, groups of 4 and 5 *competing* members, require more or less 11 and 30 iterations, respectively.

The MIL of the various simulated groups' discussions is reported in Figure 13(b). We notice that, even in this case, the MIL of uncooperative groups is higher than that of baseline and cooperative groups, particularly in large groups. The obtained results explain why in real situations groups of uncooperative people often fail to reach any consensus.

With the multiple pairwise comparisons, we find out that only *competing* and *accommodating* produce significantly different results in groups of size 2 ($p = 0.026$). In groups of size 3, *collaborating* has a significantly different MIL than *competing* ($p = 0.004$), *avoiding* ($p = 4e^{-6}$) and *baseline* ($p = 0.001$). As reported in Table 10 the significant comparisons found in groups of size 4 comprise *accommodating* and *competing*, *avoiding* and *accommodating*, *collaborating* and *competing*, *collaborating* and *avoiding*, and *baseline* and *avoiding*. This holds for groups of size 5 as well.

Table 10 The Bonferroni corrected p-values for pairwise comparisons in terms of MIL in groups of size 4 whose members have diverse interests but similar conflict resolution styles, stopping at the first group choice

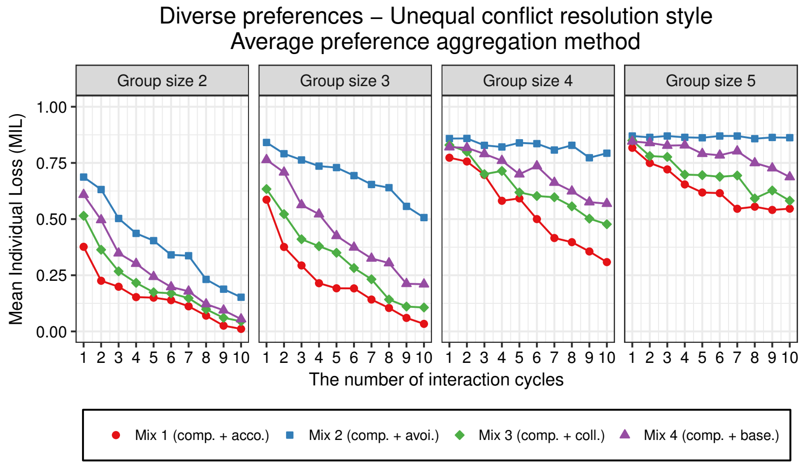
	compete	accommodate	avoid	collaborate
accommodate	0.00	-	-	-
avoid	1.00	$1.6e^{-7}$	-	-
collaborate	0.00	1.00	$8.1e^{-9}$	-
baseline	0.12	0.25	0.00	0.09

6.2.2 Mixed conflict resolution styles

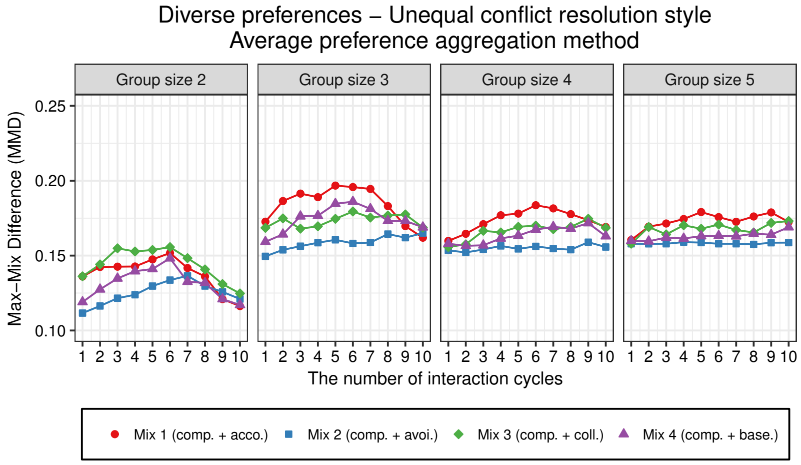
We now consider groups of users with different preferences and mixed conflict resolution styles.

Varying the interaction length. Figure 14(a) shows that the combination of *competing* and cooperative users (see Mix 1 and Mix 3) leads to the lowest loss in terms of MIL. But when the *competing* users are mixed with *avoiding* (Mix 2) or with *baseline* members (Mix 4), the average loss becomes the largest and second largest, respectively.

Figure 14(b) visualizes the utility difference between the winner and the loser in different combinations. Despite having the largest average loss, the difference between the winner and the loser's utility in mixed groups of *competing* and *avoiding* users is the lowest. MIL and MMD in mixed groups of uncooperative users are also moderately stable in larger groups (i.e., groups of size 4 and 5). Moreover, we have observed in the experimental data that most of the time *competing* members are the winners when they are paired with either *accommodating* or *collaborating* users.



(a) MIL of the mixed combinations

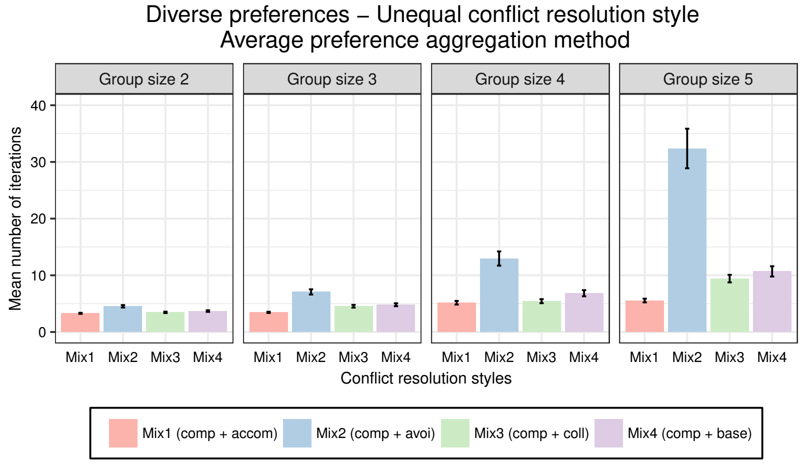


(b) MMD of the mixed combinations

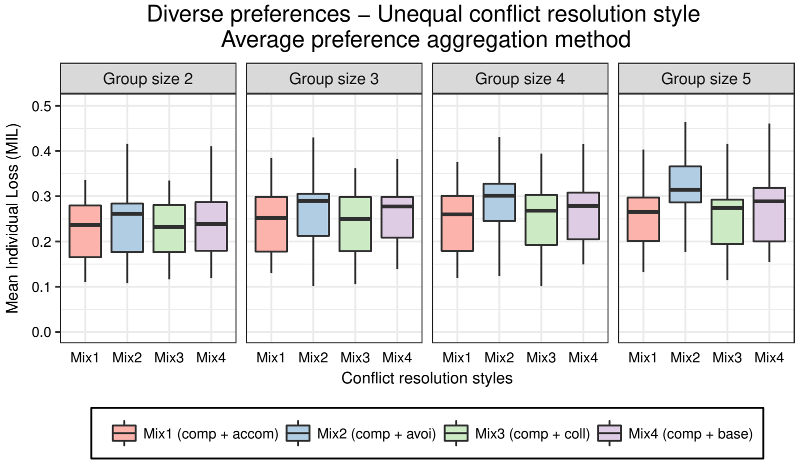
Fig. 14 Performance of groups with diverse interests and mixed conflict resolution styles

With respect to MIL and MMD, significant differences between the diverse mixtures of conflict resolution styles are confirmed by the Kruskal-Wallis test. We also find that the pairwise differences between mixtures are significant in each group size.

Stopping the discussion at the first group choice. Figure 15(a) shows the average number of interaction cycles required for identifying the group choice in each situation. In line with the previous group situations, when *competing* and *avoiding* users are paired with each other, more iterations are needed to find the first group choice. They also suffer the largest loss with re-



(a) Mean number of iterations of the mixed combinations



(b) MIL of the mixed combinations

Fig. 15 Performance of groups with diverse interests and mixed conflict resolution styles, stopping the discussion at the first group choice

spect to MIL, which is illustrated in Figure 15(b). In groups of 4 and 5 users, the mixed style groups, overall, arrive at the group choice relatively late, i.e., on average, they need 5 and 10 iterations to find the collective choice. From this observation, we can say that group size does matter for groups consisting of competitive members, hence in realistic settings they will be unlikely to reach a group decision.

Considering a 0.05 level, we find statistically significant differences of MIL in groups of size 3 ($p = 0.03$), size 4 ($p = 0.02$) and size 5 ($1.02e^{-11}$), but this difference is not found for groups of size 2 ($p = 0.24$). The mixed groups of

Table 11 Summary results of the experiment with the heterogeneous groups

Interaction length	TKI styles	MIL	MMD	# Iterations
Varying length	Cooperative groups (*collaborate* or *accommodate*)	smallest	largest	1..10
	Uncooperative groups (*compete* or *avoid*)	largest	smallest	1..10
	Mix (*compete* & *accommodate*)	smallest	largest	1..10
	Mix (*compete* & *avoid*)	largest	smallest	1..10
Stopping at the first group choice	Cooperative groups	smallest	largest	smallest
	Uncooperative groups	largest	smallest	largest
	Mix (*compete* & *accommodate*)	smallest	largest	smallest
	Mix (*compete* & *avoid*)	largest	smallest	largest

The larger the group size, the greater the effect.

competing and *accommodating* have MIL that are significantly different from the mixed ones of *competing* and *avoiding* in groups size 3, 4 and 5. Besides, there are statistical differences between the groups where *competing* members are paired with *avoiding* people and the groups where they are respectively mixed with *collaborating* and *baseline* members in groups of size 4 and 5.

The experimental results for the heterogeneous case wherein group members have diverse preferences are summarized in Table 11.

7 Analysing real groups

In this section, we are focusing on real groups and their discussions. The goal is to show that similar patterns and outcomes of the simulated group discussions can be found in real groups as well. This supports the validity of the proposed simulation approach, that is, with simulations, group behaviour can be explored in more detail without the restrictions imposed by the uncontrolled variables of real groups.

7.1 Data collection

In order to compare simulations with real groups, we have applied to real group discussions a suitable data collection process, especially designed for the travel and tourism domain. The process was organized in three phases, in which groups were observed before, during and after their discussion. The process was carried out as a part of regular lectures at TU Wien. The participation was voluntary and participants (i.e., students) were rewarded with additional points for the affiliating course. Prior to the first study phase (group discussion and decision making), the participants were briefly introduced to the experimental procedure, and they were instructed to form groups of minimum four and maximum six group members. Next, still before the first study phase, groups were asked to choose two members who would observe their group discussion. Group members who performed observations are referred as observers, and the others, who participated in group discussions, as decision-makers. Thus, each group contained at least two, and a maximum of four decision-makers.

In the first phase, the pre-discussion phase, decision-makers were asked to fill-in an online questionnaire (further referred to as the pre-questionnaire). The pre-questionnaire tapped into the individual (i.e., group-unrelated) preferences about ten pre-selected destinations (large European cities). These preferences were expressed in the form of a ranking of the ten destinations. Moreover, in this phase, observers were trained. The purpose was to present them with the details of the second and third phases, and to instruct them on how to perform and document the observations of group behaviour. A report template for documenting the group behaviour, i.e., actions of the decision-makers, based on Bales's Interaction Process Analysis (IPA) [2], was explained and distributed to the observers. IPA is a coding method for observing group interactions and it is widely used as it increases the objectivity of observations. The approach requires that an observer identify a "*unit*" of interaction for each group member. Bales defines a "*unit*" of interaction as a single simple sentence or its equivalent. For example, if a group member states "*How about voting, but I think we still might not get the winner.*", the observer should break down the sentence into two "*units*": 1)"*How about voting*", and 2) "*I think we still might not get the winner.*". Furthermore, in addition to speech, a "*unit*" of interaction includes also facial expressions, gestures, body attitudes, emotional signs, etc. Then, for each group member, the observer categorizes each "*unit*" of interaction into one among twelve behaviour categories:

1. *Show solidarity* / "*Friendly*" (e.g., expressing gratitude or appreciation; apologizing, or smiling directly at another; offering assistance, time, energy, money; etc.);
2. *Show tension release* (e.g., showing cheerfulness, satisfaction, enjoyment, relish, pleasure, etc.);
3. *Agree* (e.g., agreement reflected through verbal or nonverbal expressions);

4-6. 4. *Give suggestion* (e.g., mentioning a problem to be discussed: "*I want to call your attention to the budget issue*") / 5. *Give opinion* (e.g., stating judgement or inference: "*I believe that Amsterdam is the most beautiful place to visit in spring*") / 6. *Give information* (e.g., reporting factual, verifiable observations or experiences: "*The weather in Amsterdam at this time is not good*");

7-9. 7. *Ask for suggestion* (e.g., requesting guidance in problem-solving process) / 8. *Ask for opinion* (e.g., questions seeking value judgement, beliefs or attitudes) / 9. *Ask for information* (e.g., questions requesting a simple factual, descriptive, objective type of answer);

10. *Disagree* (e.g., rejecting another person's statement);
11. *Show tension* (e.g., appearing startled, blushing, showing embarrassment);

12. *Show antagonism* / "*Unfriendly*" (e.g., attempting to override the other in conversation, interrupting the other, making fun of others, criticizing, ill-treating, tricking, deceiving, etc.).

The observers also received detailed written explanations on how to perform observations and a continuous contact with them was maintained until the end of the data collection procedure.

In the second phase, the actual group discussions took place. The groups were introduced to a naturalistic scenario, in which they as a group should decide on a destination to visit jointly. They were also asked to choose the second best option in case the first option was unavailable. No other instructions were provided to the groups. This specific design was chosen due to its simplicity. Usually, when a group is planning a trip a number of different trip aspects have to be considered, e.g., timing, budget, destination, accommodation, transport, etc. A proper discussion on all these issues would be almost impossible to simulate in a controlled environment. Thus, we concentrated on a simple aspect, i.e., the selection of a destination, to analyse the basis of group interactions and dynamics in this specific context. The observers were included in the group work. They audio recorded and documented the group discussions using the Bales's IPA report template.

In the third, final phase of the study, decision-makers filled-in another online questionnaire (further referred to as the post-questionnaire). In the post-questionnaire, participants entered their first and second group choice, their individual choice satisfaction, difficulty of the decision making process, etc. In this analysis, we only use information about the group choice, therefore we will not further explain the other questionnaire constructs. During this phase, interviews with the observers were arranged: for each group, one meeting with the two observers took place. At the interviews, firstly, we asked the observers to explain different sections of their report template and behaviour categories in order to evaluate their understanding of the task they were given. Secondly, the two observers elaborated their own submissions and compared them, if the recordings differed to a great extent, the observers were asked to come to an agreement and revise their reports.

The data collection process resulted in 27 participants organized in eight groups, where two groups consisted of two members, one of three, and five groups of four group members.

7.2 Methodology

The first step in the analysis was to define a suitable proxy of the MIL metric that was used to assess the outcome of the simulated group discussions. Then, we had to assign each group member and group to an appropriate conflict resolution style. With respect to MIL, the user model captured by the user study differs from the one used in the simulation analysis, since the preferences are expressed as ranked lists. Therefore, to measure MIL, we adopted a different approach, while trying to identify a metric that captures essentially the same signal. First, for each group member we calculate the Spearman Footrule distance between their preferences and the group choice:

$$DIST(u,g) = rank_u(group\_choice(g)) - 1.$$

More precisely, the distance is calculated as the difference between the individual and group rank of the group choice (i.e., group rank is clearly 1, as it is the preferred group option). Then, MIL is simply calculated as the mean value of the previously defined individual group members' distances

$$MIL(g) = \frac{1}{|g|} \sum_{u \in g} DIST(u,g).$$

To assign group members, and therefore groups, to their "appropriate" TKI conflict resolution styles, we used the observational data, collected during the group discussions. Here, it was crucial to have a clearly defined procedure as in the simulation analysis. To this end, only five of the total twelve behavioural categories were used, i.e., *Friendly*, *Agree*, *Disagree*, *Opinion give*, and *Information give*. Again, the four conflict resolution styles, plus the baseline category, were defined upon two dimensions, *assertiveness* and *cooperativeness. Assertive* participants were defined as those with a high number of repetitions recorded for the *Opinion give* or *Information give* categories, while *unassertive* participants were simply those who were not identified as *assertive. Cooperative* participants were defined as those with a high number of repetitions recorded for the *Friendly* or *Agree* categories, and *uncooperative* participants as those with a high number of repetitions recorded for the *Disagree* category. The number of recorded repetitions was considered as high when the individual score was higher than the mean number of times the action was recorded over the whole data set.

Finally, a group was assigned to a specific conflict resolution style when the majority of its group members belong to that particular conflict resolution style. This particular choice was made since in reality it is almost impossible to observe "pure" groups, in terms of group members' Thomas-Kilmann conflict resolution styles. Moreover, in the simulation analysis, mixed groups have more or less equal frequency of the combined conflict resolution-styles. Therefore, it does not make sense to categorize a real group as mixed, when the majority of participants belong to a certain style (i.e., the frequency of group members' conflict resolution styles is not equal, as in the simulated groups). At the end, this resulted in eight groups with equal conflict resolution styles, of which five belonged to the baseline and three to the collaborative style. Table 12 shows repetitions of selected behavioural actions at the group level, together with the assigned TKI conflict resolution style.

7.3 Results

The results obtained by this analysis can be compared with the simulation results obtained from the simulations of group behaviour with diverse preferences

Table 12 Group discussion behaviour

Group	*Friendly*	*Agree*	*Disagree*	*Info give*	*Opinion give*	TKI style
1	3.00	3.00	2.50	2.00	4.00	baseline
2	8.00	1.50	10.50	3.50	6.50	baseline
3	19.50	14.50	9.50	27.50	20.50	baseline
4	7.50	.50	1.00	14.50	28.50	baseline
5	11.00	5.50	3.00	10.00	13.00	baseline
6	7.00	1.00	4.00	14.00	18.00	collaborative
7	9.00	1.00	5.00	10.50	8.50	collaborative
8	36.00	1.00	.00	21.50	33.00	collaborative

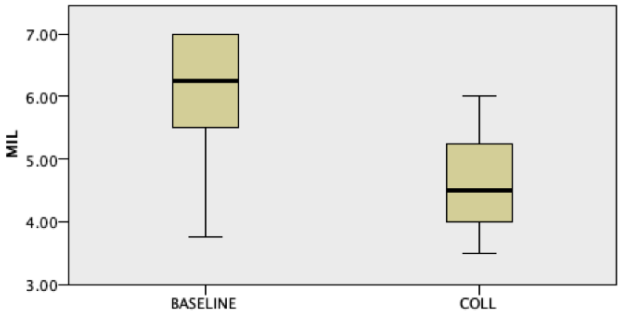


Fig. 16 MIL of real groups across uniform conflict resolution styles

and discussions that stopped with the first group choice. The reason behind this is that the diversity of group members' preferences was not controlled during the group formation phase. Moreover, measuring the group diversity with the Spearman's foot-rule distance between pairs of group members' individual top-choices, we have observed that majority of our groups are highly diverse, i.e., five of eight groups score higher than the mean diversity score measured over the set of 55 groups and 200 participants (the data set is presented in details in [15]). Finally, the groups did not receive any instructions about their discussion, therefore we can conclude that the process was finished as soon as the group found an acceptable group choice.

Hereby, we focus on the order of MIL for the two conflict resolution style categories that we obtained in the collected data set (i.e., all together seven groups, five belonging to the baseline and two to the collaborative categories). Figure 16 illustrates MIL scores for the two conflict resolution style categories, having $MIL(baseline) > MIL(collaborative)$. Unfortunately, we did not manage to capture group behaviour across all five conflict resolution style categories, however the two categories that we do have, follow the same pattern as in the simulation analysis.

7.4 Discussion

The results obtained by the analysis of real groups support a part of the findings from the simulation analysis. However, this section also clearly indicates all the difficulties, issues, and limitations that emerge by analysing real groups. Firstly, the data collection procedure is complex and time consuming. Secondly, finding participants willing to complete the three phases, and observers capable of recording group behaviour, is hard. Thirdly, even when overcoming the first two issues, it is highly unlikely to collect data that contains various conditions for conducting a thorough analysis. However, even with all its limitations, analysis of real groups is crucial to support the validity of the research results obtained by simulations. In the analysis we have presented, one clear limitation is the number of groups, and their distribution over the conflict resolution styles. Nevertheless, it does encourage further simulation analyses.

8 Conclusions and outlook

In the following, we highlight the key results of our research and improvements that can be made to the proposed simulation model.

8.1 Summary

In this paper, we have introduced a model that simulates group discussion in interactive Group Recommender Systems (GRSs). It can be used to analyze the effect of different group compositions. In the model the users' behaviour is determined by the user utility function and conflict resolution style. To assure that the model can predict realistic outcomes of the simulated discussions, it is informed by the observation of how users interact with a group recommender system. The analysis of a range of simulated group discussions has shed light on how the outcome of the group decision making process supported by the GRS is influenced by the conflict resolution styles interlinked with the inner-group similarity, interaction length and group size.

Based on the simulation results, we can state that, when group members have similar tastes, no matter what the conflict resolution styles, there is no difference in their utility for the group choice. However, groups whose members have *competing* or *avoiding* conflict resolution style get a lower utility compared to those who adopt *accommodating* and *collaborating* styles. On the other hand, in the case where the members have diverse preferences, we have found an interesting tension between the average individual's utility loss and the fairness of the system-generated group recommendations. In particular, groups of *competing* or *avoiding* users often select the recommendation that leads them to lose their utility equally to each other, but it also makes them suffer the greatest loss on average. We have observed the opposite trend in groups of *accommodating* and *collaborating* users. When it comes to groups of mixed conflict resolution styles, we noticed that when a group is formed by a combination of *competing* and either *accommodating* or *collaborating* users, the average individual's loss is the smallest, even though the discrepancy in their utility is the largest.

In addition, the preliminary analysis conducted on data collected from the observational study on real group discussions is somewhat aligned with some of the findings of the proposed simulation model. The results of real groups illustrate that the average loss, measured in terms of MIL, of groups composed of *collaborating* styles is lower than that of groups with *baseline* style, which matches the pattern observed in the simulation experiment.

8.2 Limitations and future work

This study, undoubtedly, has a number of limitations, which ultimately are linked to the simplifying assumptions that one must make in a simulation process.

The first limitation of our work is not taking into account possible social relationships between the simulated group members. As pointed out in the group recommendation literature [17, 28], the relationship strength and their personality could impact on people's judgments. Secondly, the simulation model proposed in this work is restricted to the assumption that all group members take equal roles, which is, however, not always the case. In practice, particular members tend to have specific roles in a group (e.g., parents or children). Therefore, each member's role can be brought into play to understand their decisions [1, 7]. The third restriction of the study is that we did not simulate the case wherein the agents have no loyalty to the preferences of the other agents in order to gain their desired benefit. This is an aspect of behaviour that is typically investigated in game theory. To the best of our knowledge, this has never been considered in GRSs.

In the future, we aim at testing further whether the simulated behaviours and system outcome will actually be observed when group members with known TKI are interacting with the system. Besides, we also plan to investigate the quality of the group outcome supported by the interactive GRSs with the aforementioned dimensions together with different measurement metrics. Ultimately, based on the findings of the simulation study, our next focus of interest is the development of a mechanism for the automatic optimization of the preference learning model.

References

1. Ardissono L, Goy A, Petrone G, Segnan M, Torasso P (2003) Intrigue: personalized recommendation of tourist attractions for desktop and hand held devices. Applied Artificial Intelligence 17(8-9):687–714

2. Bales RF (1950) A set of categories for the analysis of small group interaction. American Sociological Review 15:257–263
3. Baltrunas L, Makcinskas T, Ricci F (2010) Group recommendations with rank aggregation and collaborative filtering. In: Proceedings of the 4th ACM conference on Recommender systems, pp 119–126
4. Barile F, Masthoff J, Rossi S (2017) The adaptation of an individual's satisfaction to group context: the role of ties strength and conflicts. In: Proceedings of the 25th Conference on User Modeling, Adaptation and Personalization, ACM, pp 357–358
5. Bekkerman P, Kraus S, Ricci F (2006) Applying cooperative negotiation methodology to group recommendation problem. In: Proceedings of Workshop on Recommender Systems in 17th European Conference on Artificial Intelligence (ECAI 2006), pp 72–75
6. Ben-Akiva ME, Lerman SR (1985) Discrete choice analysis: theory and application to travel demand, vol 9. MIT press
7. Berkovsky S, Freyne J (2010) Group-based recipe recommendations: analysis of data aggreagation strategies. In: Proceedings of the 4th ACM conference on Recommender systems, pp 111–118
8. Blanco H, Ricci F (2013) Inferring user utility for query revision recommendation. In: Proceedings of the 28th ACM Symposium on Applied Computing, pp 245–252
9. Boratto L, Carta S (2015) The rating prediction task in a group recommender system that automatically detects groups: architectures, algorithms, and performance evaluation. Journal of Intelligent Information Systems 45(2):221–245
10. Braunhofer M, Elahi M, Ricci F, Schievenin T (2013) Context-aware points of interest suggestion with dynamic weather data management. In: Information and communication technologies in tourism 2014, pp 87–100
11. De Pessemier T, Dooms S, Martens L (2014) Comparison of group recommendation algorithms. Multimedia Tools and Applications 72(3):2497–2541
12. Delic A, Masthoff J (2018) Group recommender systems. In: Proceedings of the 26th Conference on User Modeling, Adaptation and Personalization, ACM, pp 377–378
13. Delic A, Neidhardt J, Nguyen TN, Ricci F, Rook L, Werthner H, Zanker M (2016) Observing group decision making processes. In: Proceedings of the 10th ACM Conference on Recommender Systems, pp 147–150
14. Delic A, Neidhardt J, Rook L, Werthner H, Zanker M (2017) Researching individual satisfaction with group decisions in tourism: experimental evidence. In: Information and Communication Technologies in Tourism 2017, Springer, pp 73–85
15. Delic A, Neidhardt J, Nguyen TN, Ricci F (2018) An observational user study for group recommender systems in the tourism domain. Information Technology & Tourism 19(1-4):87–116
16. Forsyth DR (2014) Group Dynamics, 6th edn. Wadsworth Cengage Learning

17. Gartrell M, Xing X, Lv Q, Beach A, Han R, Mishra S, Seada K (2010) Enhancing group recommendation by incorporating social relationship interactions. In: Proceedings of the 16th ACM international conference on Supporting group work, ACM, pp 97–106
18. Guzzi F, Ricci F, Burke R (2011) Interactive multi-party critiquing for group recommendation. In: Proceedings of the 5th ACM Conference on Recommender systems, pp 265–268
19. Jameson A (2004) More than the sum of its members: challenges for group recommender systems. In: Proceedings of the working conference on Advanced visual interfaces, pp 48–54
20. Jameson A, Smyth B (2007) Recommendation to groups. The Adaptive Web, LNCS 4321:596–627
21. Kilmann RH, Thomas KW (1977) Developing a forced-choice measure of conflict-handling behavior: The" mode" instrument. Educational and psychological measurement 37(2):309–325
22. Knijnenburg BP, Reijmer NJ, Willemsen MC (2011) Each to his own: how different users call for different interaction methods in recommender systems. In: Proceedings of the 5th ACM conference on Recommender systems, pp 141–148
23. Lops P, De Gemmis M, Semeraro G (2011) Content-based recommender systems: State of the art and trends. In: Recommender systems handbook, Springer, pp 73–105
24. Mahmood T, Ricci F (2007) Learning and adaptivity in interactive recommender systems. In: Proceedings of the 9th international conference on Electronic commerce, pp 75–84
25. Márquez JOÁ, Ziegler J (2015) Preference elicitation and negotiation in a group recommender system. In: Human-Computer Interaction, Springer, pp 20–37
26. Masthoff J (2004) Group modeling: Selecting a sequence of television items to suit a group of viewers. Personalized Digital Television pp 93–141
27. Masthoff J (2015) Group recommender systems: aggregation, satisfaction and group attributes. In: Ricci F, Rokach L, Shapira B (eds) Recommender Systems Handbook, Springer, pp 743–776
28. Masthoff J, Gatt A (2006) In pursuit of satisfaction and the prevention of embarrassment: affective state in group recommender systems. User Modeling and User-Adapted Interaction 16(3-4):281–319
29. McCarthy K, Salamó M, Coyle L, McGinty L, Smyth B, Nixon P (2006) Cats: A synchronous approach to collaborative group recommendation. In: Florida Artificial Intelligence Research Society Conference, pp 86–91
30. McGinty L, Smyth B (2002) Comparison-based recommendation. In: European Conference on Case-based Reasoning, pp 575–589
31. Nguyen TN, Ricci F (2017) Dynamic elicitation of user preferences in a chat-based group recommender system. In: Proceedings of the 32nd ACM Symposium on Applied Computing, pp 1685–1692
32. Nguyen TN, Ricci F (2018) A chat-based group recommender system for tourism. Information Technology & Tourism 18(1):5–28

33. Nguyen TN, Ricci F (2018) Situation-dependent combination of long-term and session-based preferences in group recommendations: an experimental analysis. In: Proceedings of the 33rd Annual ACM Symposium on Applied Computing, ACM, pp 1366–1373
34. Osogami T (2017) Human choice and good choice. In: The Role and Importance of Mathematics in Innovation, Springer, pp 1–10
35. Quijano-Sanchez L, Recio-Garcia JA, Diaz-Agudo B, Jimenez-Diaz G (2013) Social factors in group recommender systems. ACM Transactions on Intelligent Systems and Technology (TIST) 4(1):8
36. Recio-Garcia JA, Jimenez-Diaz G, Sanchez-Ruiz AA, Diaz-Agudo B (2009) Personality aware recommendations to groups. In: Proceedings of the third ACM conference on Recommender systems, ACM, pp 325–328
37. Ricci F, Rokach L, Shapira B (2015) Recommender systems: introduction and challenges. In: Recommender Systems Handbook, Springer US, pp 1–34
38. Rosenfeld A, Kraus S (2018) Predicting human decision-making: From prediction to action. Synthesis Lectures on Artificial Intelligence and Machine Learning 12(1):1–150
39. Rossi S, Di Napoli C, Barile F, Liguori L (2016) A multi-agent system for group decision support based on conflict resolution styles. In: International Workshop on Conflict Resolution in Decision Making, pp 134–148
40. Stettinger M, Felfernig A, Leitner G, Reiterer S, Jeran M (2015) Counteracting serial position effects in the choicla group decision support environment. In: Proceedings of the 20th International Conference on Intelligent User Interfaces, pp 148–157
41. Thomas KW (2008) Thomas-kilmann conflict mode. TKI Profile and Interpretive Report pp 1–11
42. Tkalcic M, Delic A, Felfernig A (2018) Personality, emotions, and group dynamics. Springer
43. Trabelsi W, Wilson N, Bridge D, Ricci F (2010) Comparing approaches to preference dominance for conversational recommenders. In: Proceedings of the 22nd IEEE International Conference on Tools with Artificial Intelligence, pp 113–120
44. Trattner C, Said A, Boratto L, Felfernig A (2018) Evaluating group recommender systems. Springer
45. Viappiani P, Pu P, Faltings B (2008) Preference-based search with adaptive recommendations. AI Communications 21(2-3):155–175
46. Wood VF, Bell PA (2008) Predicting interpersonal conflict resolution styles from personality characteristics. Personality and Individual Differences 45(2):126–131

Author Biographies

Thuy Ngoc Nguyen is currently a research assistant at the Faculty of Computer Science at the Free University of Bozen-Bolzano, Italy, where she received her PhD for studies on supporting group discussions with recommendation techniques. Her main research interests include human computer interaction, recommender systems, and user modeling.

Francesco Ricci is a professor of Computer Science at the Free University of Bozen-Bolzano, Italy. His research interests include recommender systems, intelligent interfaces, machine learning and the applications of ICT to health and tourism. He has published more than one hundred fifty of academic papers on these topics. He is on the editorial board of User Modeling and User Adapted Interaction, the Journal of Information Technology & Tourism and CCF Transactions on Pervasive Computing and Interaction.

Amra Delic is a final year PhD student at the Faculty of Informatics in Vienna. Her PhD research deals with group recommender systems in the travel and tourism domain. Specifically, her research covers a variety of human behavior aspects, such as personality traits, travel behavioral patterns, intra-group social relationships, social identity theory, choice satisfaction, etc., providing a more in-depth, and comprehensive overview of group modeling for group recommender systems.

Derek Bridge is a senior lecturer in the Insight Centre for Data Analytics based in the School of Computer Science and Information Technology at University College Cork, Ireland. At national level within the Insight Centre, he leads the Recommender Systems Group.