A New Approach to Large-Scale Deliberation

Sanja Tanasijevic
Karlsruhe Institute of Technology (KIT), Germany
sannya.tanasijevic@kit.edu

Klemens Böhm
Karlsruhe Institute of Technology (KIT), Germany
klemens.boehm@kit.edu

***Abstract*— In this article we propose a novel approach for group-decision making, i.e., proposing and selecting solutions to issues discussed in online settings, based on the structure of the discussion. Our approach consists of three steps: (1) assigning weights to participants based on formal criteria such as degree of engagement in the discussion; (2) assigning scores to comments, considering the weights of authors and raters; (3) assigning scores to proposals, based on the scores of the pro and contra arguments. An important point is that individuals whose behavior is in line with our formal criteria have a higher influence on the decisions. Having built a respective online platform, we have evaluated the proposed model by means of an experiment with more than 100 participants who have discussed several topics relevant to them and a subsequent survey. Looking at both the results of survey and the discussion outcomes, we conclude that our approach yields comprehensive discussions and outcome decisions supported by the community.**

Keywords—deliberation; online forum; group decision-making

I. Introduction

The question how communities can come to decisions and solutions continues to be fundamentally important. Online discussions sometimes generate poorly organized, unsystematic and redundant contributions of varying quality [1]. Significant effort is required to extract important issues, ideas and arguments. This includes discussions on budget planning for the German city of Essen (essen-kriegt-die-kurve.de), to give an example. Its limitations are exemplary of the ones of many other initiatives. The project lets individuals propose concrete budget cuts and discuss these proposals. It also tries to come to some conclusions from the discussion, by using rather simple quantitative measures such as the number of pro arguments regarding a proposal. However, this does not say much about the importance and relevance of the various arguments. In particular, people have started discussing issues not related to the proposal and have repeated arguments; this has affected those measures nevertheless. To address the various challenges, it is mandatory to radically reduce redundancy and encourage clarity. In a nutshell, the question investigated here is how online deliberation, i.e., the thoughtful consideration of all sides of an issue, can be facilitated so that important ideas and arguments are indeed identified, and group-decision-making is efficient. While typical papers on social network analysis study characteristics of the network structure, our current research focuses on the question how frame conditions affect this structure (the structure of the discussion in our case and user satisfaction with it).

Designing a platform that allows deriving decisions from the discussion is challenging. In real life and in other studies, e.g., ConsiderIt [2], Deliberatorium [3], one usually takes pro and contra arguments into account when making a decision. We do the same, for each proposal: Proposals are discussed in different threads where people can provide pro and contra arguments for each one, and an automated scheme selects a winner proposal in the end. In other words, decision-making is mainly based on the structure of the arguments, as opposed to voting. Thus, a first challenge is to decide which information to collect from individuals. At first sight, information that is useful includes whether an individual agrees or disagrees with a comment, or feedback on which comments he deems off-topic, repetitions etc. However, we need to flesh out which information is indeed collected. A subsequent challenge then is how we use this information to come to a decision. The decision-making scheme must be understandable and non-ambiguous, cf. [4] [5]. Finally, evaluating any approach that claims to foster deliberation is challenging as well.

Our contributions are as follows: First, we motivate and propose various criteria that constitute desirable behavior of community members, e.g., originality of arguments, focus on the topic in question etc., and propose formalizations of each of them. To stimulate desirable behavior, each community member has a weight that depends on the degree of adherence to our criteria. The weight determines his influence on the decision to be taken. Next, we propose a decision-making scheme that is argument-based. With our scheme, each argument is assigned a score that depends on the degree of agreement it has obtained from the community and on the weights of the respective individuals. We also formalize when an argument is rejected, i.e., ignored by the decision-making scheme. Our scheme assigns each proposal a score that depends on the share of pro and contra arguments and their scores. In our setup, proposals are alternatives to each other, and the proposal with the highest score will be the winner proposal. Finally, we evaluate our approach in a setting that is very close to a real one, with more than 100 participants. Students of a database course at our university have deliberated on various topics relevant to them. While we have not been able to validate all our hypotheses in our specific setup, an important result is that a majority has expressed satisfaction with the weighting criteria and the decision-making scheme, and they have given preference to our forum model over plain voting in terms of quality of decisions taken, mutual respect of opinion etc.

II. Related Work

Deliberation is a process where communities (1) identify possible solutions for a problem, and (2) select the solution(s) from this space that best meet their needs [6] [7]. Its strengths are idea synergy and diversity, results checked by many and collective wisdom. Deliberation is touted as a form of discussion where participants share their considerations in order to make decisions of higher quality and legitimacy [8] [9] [10] [11] [12] [13] [14] [15]. In practice however, deliberation faces serious challenges, including disorganized, redundant content, quantity over depth, strong polarization, and dysfunction arguments [1]. Large-scale argumentation systems claim to address these shortcomings, by providing a systematic structure that radically reduces redundancy and encourages clarity [1]. Respective projects structure content by means of different argumentation models. Thus, deliberation, in so-called argument map contexts in particular, tends to evolve from defining issues to proposing ideas to identifying trees of pro and con arguments [3]. The Deliberatorium project has used the IBIS argumentation formalism [16]. There, members of a community make their contributions in the form of a deliberation map, a tree-structured network of posts each representing a single unique issue (question to be answered), ideas (possible answers to a question), or arguments (pros or cons for an idea or another argument) [3]. The process of building a tree is under moderator control. Similarly, the Cohere project has aimed to establish a tool for distributed and asynchronous argumentation [17]. Default roles in the project are Questions, Answers, Pros and Cons, derived from IBIS. Visualization of argumentation schemes and critical questions, as proposed by Walton [18], can also be modeled in Cohere. The IBIS formalism and other ones have limited application in real-life scenarios, due to acceptance of discourse and classification problems related to the completeness, comprehensiveness and pedantry of the classification [16]. We for our part have targeted at an argumentation model that is intuitive rather than exhaustive. Similarly, in a rather informal fashion, ConsiderIt [2] has proposed content structuring. It has encouraged discussants to deliberate by formulating pro/con points, which participants can create, further share and adopt. But in contrast to our approach, ConsiderIt does not feature decision-making. Slashdot also has gone far in its efforts to structure online conversations and has implemented distributed moderation [19]. Actual participants do the work, and their influence depends on their reputation in the community. The nature of the conversations is different from ours and does not call for group decisions, so Slashdot's ranking and scoring schemes are not applicable in our context.

Most of the argument-based approaches presented above aim to not only structure the content, but also facilitate recognizing important points and arguments and finally select a decision. The E-liberate project has based decision-making in communities in equitable and collective manner by employing Robert's rules of Order [20]. Different roles are supported: chairs, members and observers. Content classification is text-based. However, text recognition and text-based classification still have issues, and strict rules on participation and presence in the discussion cannot be really enforced in an online environment.

A related research question is expanding perspectives on the issues in question. The NewsCube project has tried to broaden views on news by giving several viewpoints [21]. Reflect is a system that engages and motivates discussants to restate, identify and share common grounds [22]. The background of this project is the listening part of deliberative considerations, which is important in the community discussions and decision-making. With our approach, participants are incentivized to ‚listen‘ as well, since they will only have a high rank if their posts are free of repetitions. Opinion Space is an online interface incorporating ideas of deliberative polling, collaborative filtering for visualization and navigation through diverse comments [23]. The tool resulting from the Cohere project tries to establish a system of social networking and reputation in the community through idea linking.

III. Deliberation forum model

We now describe our forum model in detail. By intention, the look-and-feel of our forum is the one of a conventional forum wherever possible. This is because user acceptance is crucial in our context. Further, we can leverage existing technology and, hence, the host of comfort features provided by current implementations. – We will now elaborate on the discussion structure, the argumentation model, and the implementation.

A. *Discussion structure and comments types*

The following is a comprehensive description of the discussion structure and its representation in our model. The discussion structure has the following elements:

Forum (issue). A forum corresponds to the subject of discussion, e.g., “How should EUR 500 be spent?”.

Thread. Each thread within its forum discusses one specific suggestion on how the issue in question could be solved.

Comments. Comments are the constituents of a thread, i.e., a comment is always part of a specific thread. Comments are typed, e.g., pro argument or contra argument. A comment can refer to another comment.

Ratings. A rating expresses the perspective of an individual on a comment posted by someone else. In our context, a rating is a complex structure consisting of various attributes, e.g., whether the individual agrees or disagrees with the comment, how he evaluates its writing style or its tone etc.

There are different comment types, mimicking common argumentation structures:

A proposal is a suggestion how to solve a forum issue. To illustrate, one issue in our study has been which criterion should be used to decide whom to give an iPad to. One proposal has been to give it to the student with the highest number of points in the exercise in the current semester.

Extension of a proposal. Individuals can extend a proposal by means of a comment (in contrast to issuing a new proposal). To illustrate, an extension of the proposal just mentioned has been to use the number of points in the exercise to assign a certain number of lots to individuals, and more points increase the number of lots and the probability of winning the iPad.

A **pro argument** is a comment in favor of a proposal.

A **contra argument** is a comment against a proposal.

Other is a comment which the author does not want to classify as one of the types just mentioned.

A design decision of ours has been to devote more importance to the simplicity of the model, compared to exactness and comprehensiveness. As pointed out in the previous section, literature has proposed various argumentation schemes with much sophistication. However, instead of having a model that is comprehensive but might be overly complicated for non-experts, we have limited our model to elementary comment and rating types.

B. Rating model

Participants can express their opinion on comments by others by means of ratings. In our context, a rating consists of the following attributes:

Content. Individuals can assess a comment by content using one of the following options: agreement, disagreement, repetition, and off-topic.

Writing style. Writing style can be evaluated using the grading scale from 1 to 5. Rate (5) represents clear and concise, writing style as opposed to unclear, incomplete text (1).

Tone. Analogously, tone can be (5) balanced and polite, as opposed to provocative and offensive (1).

Comment type. To ensure that comment types as specified by the authors are correct, other users can state the type of a comment as part of a rating as well. The possible values are proposal extension, pro argument, contra argument, and other.

C. Weighting scheme

To motivate community members to avoid redundancy and to contribute to clarity when deliberating, each member is assigned a weight. The weight quantifies the degree of compliance with various criteria that are helpful in our context. A high weight gives individuals higher influence on the decisions taken eventually. We now list those criteria before giving the formal definitions of some of them. Due to lack of space, please see an extended version of this article [24] for the remaining ones.

Originality. This indicator has a high value if few comments issued by the participant in question are rated as repetitions by many others.

Focus. The fewer comments by the participant are rated as off-topic, the higher will be the value of the indicator.

Style. The value of this indicator directly depends on the writing-style ratings of her/his comments.

Tone. The value of the tone indicator directly depends on the tone ratings of the comments by the participants.

Engagement. This indicator comprises the number of comments and ratings issued by the participant.

Individuality. The rationale behind this criterion is to make collusion attacks and team-ups of individuals more difficult and to curb the influence of herding behavior. Individuality is the share of participants whom the participant in question agrees with in some context and disagrees with in some other context. To illustrate, a participant being a perfect match with many other participants regarding comments and ratings has a low value regarding this criterion.

Breadth. We postulate that participants engaged in many discussion threads should be rewarded. The rationale is to curb the influence of participants with vested interests who only put attention to their specific issue.

Honesty. The rationale here is to ensure honest behavior of participants. In recent years, economic literature has proposed a number of methods to maximize the reward for individuals answering questions truthfully, even in the absence of an objective truth criterion, so-called *honest feedback mechanisms (HFM)*. For instance, the so-called *peer-prediction method* applies scoring rules to the posterior belief on ratings by others, and honest reporting turns out to be a Nash Equilibrium [25]. We for our part use the peer-prediction method; it assigns scores for each rating based on its probability compared to the reference rating [26].

While this is the list of criteria we have come up with after lengthy considerations, we do not claim at this point to have indeed covered all aspects of desirable behavior. However, we are confident not to face major difficulties when having to implement further criteria, redefining ours or even omitting some. According to our design, to not discriminate against minority opinions, the weight of an individual does not depend on the degree of agreement of the community with his arguments.

D. Formulae and notation

P is the set of all participants. $K^{create}(j)$ is the set of all comments Participant *j* has posted. *K* is the set of all comments posted in all forums. *T* is the set of all threads, and *K(t)* is the set of comments in Thread *t*. $K^{create}(j,t)$ contains the comments posted by Participant *j* in t. $F \subset T$ represents a forum. $K(F) = \bigcup_{t \in F} K(t)$ is the set of all comments in *F*. $R^{create}(j)$ is the set of ratings which Participant *j* has posted. A rating consists of the following information: content rating, writing style and tone rating, type of the comment and the rater. The type of ‘content’ is the enumeration that takes values from {agree, disagree, off-topic, repetition}. Writing style and writing tone can take values from 1 to 5. The type of ‘comment type’ is the enumeration with the following values: extension of a proposal, pro argumentation, contra argumentation and other. Each rater can submit only one rating of a comment. *R* is a set of all ratings, irrespective of who has issued them. $R(k)$ is the set of ratings on Comment k, $R^{subject}(j)$ is the set of ratings on comments issued by Participant *j*, while $R^{subject}_{off-topic}(j)$ is the set of ’off-topic‘ ratings of comments of Participant *j*.

Definition. $T^{create}(j)$ *is the set of all threads Participant j has actively participated in by posting a relatively high number of comments.*

$$T^{create}(j) := \left\{ t \in T \mid \left| K^{create}(j,t) \right| > \underset{i \in P}{avg}\left(\left| K^{create}(i,t) \right| \right) / 2 \right\}$$

Here, we only count threads where the participant has at least posted half of the average number of comments. The rationale has been to have a certain level of engagement as a prerequisite for active participation, as are other parameter values that follow.

Definition: *Breadth of Participant j.*

$$breadth(j) := \frac{\left|T^{create}(j)\right|}{|T|}$$

Definition: *Focus of Participant j.*

$$focus(j) := 1 - \frac{\left|R^{subject}_{off-topic}(j)\right|}{\left|R^{subject}(j)\right|}$$

Definition: *Originality of Participant j.*

$$orig(j) := 1 - \frac{\left|R^{subject}_{repetition}(j)\right|}{\left|R^{subject}(j)\right|}$$

Definition: *A comment is useful when less than 50% of its ratings are 'off-topic' and 'repetition'.* The set of useful comments posted by Participant *j* is $K^{create}_{useful}(j)$, while the set of all these comments unrelated to a specific author is K_{useful}.

Definition: *Engagement of Participant j.*

$$engage(j) := \frac{\left|K^{create}_{useful}(j)\right|}{\underset{i \in P}{avg}\left(\left|K^{create}_{useful}(i)\right|\right)} + \alpha_{engage} \cdot \frac{\left|R^{create}(j)\right|}{\underset{i \in P}{avg}\left(\left|R^{create}(i)\right|\right)}$$

Weight α_{engage} gives different weights to comments and ratings. Since writing a comment requires more effort than submitting a rating, we have set the ponder to 0.25 haphazardly.

Definition: *A tone rating is bad when a tone attribute has a value of 1 or 2.* $R^{create}_{tone-}(j)$ is the set of bad tone ratings of the comments that Participant *j* has posted.

Definition: *Tone of Participant j.*

$$tone(j) := 1 - \frac{\left|R^{subject}_{tone-}(j)\right|}{\left|R^{subject}(j)\right|}$$

Definition: *A writing style rating is bad if it has a value of 1 or 2.* $R^{create}_{style-}(j)$ is the set of bad style ratings of the comments Participant *j* has posted.

Definition: Writing style of Participant *j*.

$$style(j) := 1 - \frac{\left|R^{subject}_{style-}(j)\right|}{\left|R^{subject}(j)\right|}$$

indiv(j) is the individuality of Participant j, as defined in the technical report [24], *hfmscore(j)* denotes his honesty.

All indicator values are in the range [0, 1]. We have seen two alternatives to normalize these values. Here, normalization does not only take the values, but also their distribution in the community into account. The normalized value of an indicator is the share of participants who have an indicator value lower than the one of the current participant. To illustrate, if only 20% of the community have performed better than Participant *j* regarding criterion breadth, *j*'s normalized value of indicator breadth is 0.8. The advantage of this kind of normalization is that it distributes the participants over the entire [0, 1] range and makes criteria comparable. The disadvantage is when everybody performs similarly. Then slight deviations can have a significant effect. This is why we have not normalized indicators focus, originality, and style in this way. We have assumed that only a few participants would post off-topic or repetition comments, and if someone has a value slightly worse than average, this kind of normalization would have really set him back. The remaining indicators however are normalized in this way.

Definition: *Normalization of an indicator by frequency distribution*. The normalized value of an indicator of Participant *j* is the share of participants whose indicator value is less than or equal to the value of *j*. We use the notation $indicator^{norm}$, e.g., $indiv^{norm}(j)$, for normalized indicator values.

Definition: *Weight of a participant.*

$$WEIGHT(j) := \min\begin{pmatrix} focus(j), orig(j), style(j), tone(j), breadth^{norm}(j), \\ engage^{norm}(j), indiv^{norm}(j), hfmscore^{norm}(j) \end{pmatrix}$$

A participant must perform well regarding all criteria in order to have a high weight. We use the minimum function here so that this becomes clear to the user as well. It should now be obvious to him which aspects of his behavior he needs to devote more attention to in order to receive a higher weight.

E. Decision-making scheme

Our decision-making scheme is argument-based. Each argument receives a score dependent on the degree of agreement obtained from the community and the weights of the respective individuals. Next, our scheme assigns each proposal a score that depends on the pro and contra arguments and their scores. In our setup, proposals are alternatives to each other, and the one with the highest score will be the winner proposal.

1) Formulae and notation

$K_{ref}(p)$ is the set of comments in the thread belonging to Proposal *p*. $K^{+}_{ref}(p)$ is the set of pro arguments related to *p*, $K^{-}_{ref}(p)$ the set of contra arguments. The author of Comment *k* is denoted by *author(k)*. $R_{ref}(k)$ is the set of all ratings of Comment *k*. $R^{+}_{ref}(k)$ is the set of ratings of type 'agreement' while $R^{-}_{ref}(k)$ is the set of 'disagreement' ratings for *k*.

Definition: *Comment score.*

$$score(k) := \left(\frac{weight(author(k)) + \sum_{r \in R^{+}_{ref}(k)} weight(issuer(r))}{weight(author(k)) + \sum_{r \in \left(R^{+}_{ref}(k) \cup R^{-}_{ref}(k)\right)} weight(issuer(r))} - 0{,}5 \right) \cdot w_1(k)$$

$$w_1(k) =: \frac{weight(author(k)) + \sum_{r \in \left(R^+_{ref}(k) \cup R^-_{ref}(k)\right)} weight(issuer(r))}{\max_{k' \in K(F)} \left(weight(author(k')) + \sum_{r \in \left(R^+_{ref}(k) \cup R^-_{ref}(k)\right)} weight(issuer(r)) \right)}$$

The score of a comment depends on the weight of its author and raters, and on the share of agreement ratings in the set of all ratings it has received. In addition, Weight w_1 takes into account the number of participants having issued ratings of Comment *k* and normalizes the scores in the forum thread, using the maximum sum of weights of author and raters.

Definition: *Proposal score, pscore.*

$$pscore(p) = \frac{\sum_{k \in \left(K^+_{ref}(p) \cup \{p\}\right)} score_k - \sum_{k \in K^-_{ref}(p)} score_k}{\max_{p' \in F} \left(\left| \sum_{k \in \left(K^+_{ref}(p') \cup \{p'\}\right)} score_k - \sum_{k \in K^-_{ref}(p')} score_k \right| \right)}$$

The score of a proposal depends on the scores of its pro and contra arguments. The more pro arguments there are, and the higher their scores are, the higher is the proposal score. Scores are normalized on the forum level, to make scores in different forums comparable. Finally, note that individuals with low weights can still influence the outcome by coming up with proposals or arguments that a majority is in favor of.

The evaluation of proposal extensions is a difficult issue since the context of these extensions is not bounded in any way. In particular, extensions can address different perspectives of the proposal; they can mutually exclude each other or not. We have left the question how to score them as future work and have evaluated them by hand in this current study.

F. Anonymity

Our forum is anonymous. The names of authors or raters of comments are not visible. The rationale has been to indeed put the focus on the comments and the argumentation and not on the persons involved. Further, the type of a comment as specified by its author is not displayed. For instance, if a person is strongly in favor of a certain proposal, he might rate the contra arguments negative a priori without even bothering to read. Similarly, summaries of ratings of comments issued so far are not shown either to avoid influencing participants.

IV. Hypotheses

We have evaluated our forum model by means of an extensive user study. Before describing it, we list some of our hypotheses, together with their rationale. See [24] for the full list.

H1: Participants have deemed our weighting scheme fair. We are interested in the perception of the fairness of the model by participants, including our choice of criteria and the technical details of the indicator calculation.

H2: The perception of usefulness of decision-making scheme is positively correlated with the perceived fairness of the weighting model. The fairer the weights are perceived, the better is the evaluation of the decision-making scheme.

H3: The perceived fairness of the weighting scheme is positively correlated with the degree of respect for the opinions of others. This hypothesis evaluates the effects of the weighting scheme on the evaluation of proposed solutions to the discussed issues. By assigning weights to the participants, they have different degrees of influence on the decisions.

H4: The higher the perceived usefulness of decision-making scheme, the more satisfied is the community with the winner proposals. If participants perceive the decision-making scheme as useful, it should have a positive effect on their attitude towards winner proposals.

H5: The higher the evaluation of the decision-making scheme, the higher is the perceived quality of the decisions. This claim is similar to Hypothesis H4, but with the distinction that the perceived quality of the decisions is affected.

H6: The perceived quality of the decisions is positively correlated with the participants' feeling that their opinion is respected. If participants think that their opinion is respected in the community, this should affect their evaluation of the quality of decisions in a positive way.

V. Experimental setup

Our implementation of the forum model proposed so far is based on the open-source forum software *phpBB* [27]. It is written in php and uses the MySQL database for persistent data storage. *phpBB* is listed in relevant blogs as one of the top ten open-source forum projects. We have extended the existing *phpBB* platform with the specifics of our model: comment types, ratings and weighting and decision-making scheme [28]. Additionally, we have adapted the interface in order to anonymize the data, i.e., to not display information such as the names of comment authors.

We have consciously decided to evaluate our proposed model experimentally. An alternative to experiments would have been a formal analysis or simulations. A difficulty with these alleged alternatives – at this stage of the project – is that they require various assumptions, e.g., how the number of arguments generated by different individuals is distributed, what is the ratio of off-topic arguments etc.

Our experiment has had 250 participants. They were students in the database course in the fourth semester of the KIT Bachelor program in computer science. The experiment was running for four weeks. In this time period, students have discussed several issues relevant to them. To illustrate, some of the topics discussed are as follows:

How should EUR 500 be spent on behalf of the students? Only proposals in line with the German regulations on how public money may be spent are admitted by the moderator. 'Beer' is an example of a proposal that is not acceptable.

We have procured a new iPad (3rd generation, Wi-Fi, 16 GB) to give away; who should receive it. Proposals containing the names of individuals or circumscriptions of concrete individuals are not accepted, only abstract specifications such as 'the best student in the class'.

What should be the topic of a new course in the area of databases/information systems in the next academic year? We have promised that the winner proposal will indeed materialize.

For the full list of topics please see [24].

We point out that we have announced that the decisions by the group are binding to us. For instance, we have promised that we will indeed offer the course with the highest degree of agreement in the subsequent academic year (analogously with the iPad or the EUR 500).

To illustrate the effects of moderation, we have discarded the suggestion that EUR 500 should be used to buy cake to throw at each other. However, once a proposal had been approved, we have not filtered any arguments referring to it. For all issues, we have made it clear that there will only be one winner proposal, e.g., the EUR 500 will not be split. The rationale has been that we indeed wanted to study how the community deals with the situation where proposals compete with each other.

In our specific setup, a further incentive for taking part in the forum discussions were bonus points for the final exam, as follows: A participant must have posted 5 comments, none of them off-topic or repetition, and 20 ratings in order to receive a bonus of 5% of the points one could earn in the exam. With fewer comments and ratings, the bonus has been proportionally smaller. Obviously, an urgent question now is whether this bonus is the only rationale for participation. However, statements in the questionnaire and participation statistics indicate that a significant number of participants have been interested in the forum discussions themselves. Out of 163 participants who have posted at least one comment, 74 have posted more than five comments; out of 156 participants who have submitted at least one rating, 103 participants have generated more than 20 ratings. Thus, while that bonus might have influenced participant behavior, it obviously is not the only stimulus for participation.

We have decided to evaluate our forum model by means of a questionnaire [29]. At an early stage of the project, we had considered forming a committee of experts who would assess the various proposals. However, it is difficult to impossible to decide which proposal actually is good, and which one is not. To illustrate, even 'beer' might actually be a good proposal, since it fosters socializing within that community – even though the organizers of this experiment might not like it. Further, our research question is how to arrive at decisions supported by most community members after careful deliberation. Next, we point out that privacy is valued highly in Germany, and we have done the evaluation anonymously (and actually had to go through significant effort to facilitate that bonus-point regulation). In consequence, we could not relate questionnaire answers to user behavior in our system. We do plan to analyze the user data collected from our system in detail, but such a study exceeds the scope of this article.

VI. Results

In total, 250 participants have registered. 163 of them have generated at least one comment, and 156 have issued at least one rating. 116 participants have filled out the questionnaire. As described earlier, there have been seven different forum issues, and participants could generate proposals for six of them. The moderator had approved 88 proposals altogether, and 963 comments were generated in total.

We now say which hypotheses we have been able to validate in our setting.

H1: Participants have deemed our weighting scheme fair. Looking at the absolute numbers, 19 participants out of the total number of 116 participants have rated the fairness of the model as moderate. Recall that the grading scale ranges from 1 (not fair at all) to 5 (very fair). Here, 'moderate' means Rates 1 or 2. Thus, the hypothesis is confirmed. – The criteria with the highest correlation with the perceived fairness are: focus (7 moderate out of 116), tone (16 moderate out of 116), and honesty (15 out of 116). The highest positive correlation between the perception of the fairness of the weighting scheme and the fairness of the criteria is observed for the following criteria: tone ($r = 0.3566$, $p < 0.001$), individuality ($r = 0.3491$, $p <0.001$), originality ($r = 0.3357$, $p < 0.001$).

H2: The perception of the usefulness of the decision-making scheme is positively correlated with the perceived fairness of the weighting model. In the questionnaire data, there is a significant correlation ($r = 0.5433$, $p < 0.001$). In absolute numbers, only 11 participants out of 116 have rated the decision-making scheme as moderate, 32 were neutral.

H3: The perceived fairness of the weighting scheme is positively correlated with the degree of respect for the opinions of others. We have not observed a significant correlation. One possible explanation is that participants have not seen/understood how their weights affect comment scores and the evaluation of suggested solutions.

H4: The higher the perceived usefulness of decision-making scheme, the more satisfied is the community with the winner proposals. We have not observed a significant correlation that confirms this relationship. Leaving aside that we have not been able to confirm that correlation, the usefulness of the decision-making scheme is high: 11 participants out of 116 have rated the decision-making scheme as moderate, 32 were neutral. Furthermore, there is evidence that people think that their opinion is respected. Out of 114 participants who have answered the question on the respect of opinion in the forum, 73 participants have given high rates, 24 were neutral and only 11 participants have found it unsatisfactory.

H5: The higher the evaluation of the decision-making scheme, the higher is the perceived quality of the decisions. There is a certain correlation ($r = 0.2019$, $p < 0.05$), which leaves some uncertainty from a statistics point of view.

H6: The perceived quality of the decisions is positively correlated with the participants' feeling that their opinion is respected. The correlation is significant ($r = 0.2327$, $p < 0.02$). The quality of the final decision is closely related to the perceived respect of the opinion of others in the forum.

A further point is that participants were honest when rating contributions of others. Out of 116 participants 110 claimed that they had behaved honestly. Additionally, in the control question more than 65% of participants have estimated that more than 70% of participants had behaved honestly. In our opinion, such a high percentage of participants deeming a rather large group of other participants honest in many situations is a positive result. The correlation between self-reported honesty and the perceived honesty of others is significant (r = 0.3601, p < 0.001).

VII. Discussion

A. Questionnaire results

Although the questionnaire results have been helpful to answer some of our questions, there are some results that leave room for interpretation. Looking at the free-text answers, we for our part have gained the impression that the judgments on some points were sometimes based on superficial interpretations rather than on a thorough understanding of the issues. For instance, according to our web-access statistics, most participants have not fully read the documentation of the weighting and the decision-making scheme. E.g., participants have evaluated criterion ”honesty“ highly, although most of them have not been familiar with the peer-prediction method. An important insight is that, while participants have been positive about some aspects of our approach, we could not confirm all of our expectations in this specific setup. In particular, according to the questionnaire, participants have not given much attention to the weights assigned to them, and we could not validate the hypothesis that weights have affected the behavior. One possible explanation is that – in our setting – participants might have been more interested in the bonus points (which did not depend on the weights) rather than the decisions themselves. Further experiments in other settings are needed to clarify this issue. Additionally, even though the ratings of the decision-making scheme have been high, based on the questionnaire results we have not been able to confirm that, in our setting, the community was more tolerant regarding the decisions. The participants have acknowledged that respect of opinions of others is higher, and that decisions are of higher quality, but not higher tolerance towards decisions taken. Unfortunately, participants have hardly given answers to the freetext question why they have not been satisfied if this was the case.

B. Democratic principle

Clearly, weighting participants based on their behavior means that participants have different influence on the decisions. The advantage is that this should serve as an incentive to take part in the deliberation in a constructive fashion. Our approach does not violate the principle of equality according to the German constitution since it treats all participants equally. Further, our perspective is that the criteria are clear and well-documented. One's opinion does not affect the weight since our criteria are purely formal and do not include the degree of agreement/disagreement of the community with the arguments.

C. Forum model

The motivation behind our work has been to foster deliberation and to give way to decisions widely supported by the community. We have conducted our evaluation with the audience of a university course. This has some differences to other communities: First, a cohort of university students, being roughly of the same age and sharing similar academic interests, is a relatively homogeneous group of individuals, compared to other settings such as public or political discussions. Second, while bonus points have been an incentive in our context, students have shown interest in the topics discussed, i.e., two third of the students who have posted at least one rating have posted more ratings than required to receive the full bonus. However, bonus points have certainly been a stimulus for participation, and our rules for earning them have affected the behavior of participants. They must have posted a certain number of comments and ratings in order to earn this reward. These settings have advantages and disadvantages. While it might seem at first sight that this lets our approach appear in a better light, this is not necessarily the case. In particular, individuals who have only been interested in the bonus, but not in the issues to be deliberated had to generate comments and ratings. One would expect this ‘noise’ to curb the satisfaction of the rest of the community with our approach. Nevertheless, there have not been any signs of dissatisfaction. This gives way to the expectation that our approach will also work in settings without any external incentives such as bonus points. Investigating this is future work.

Another issue is that the system is to some extent vulnerable to attacks such as the following ones: Individuals can team up, earn high weights by deliberating issues of little interest to them, and then use their weights to influence decisions relevant to them. However, our criterion ’breadth’, while not ruling out this attack completely, does make it more difficult. Further, while it does not mean that this behavior pattern does not occur, participants in our study have not observed this kind of attack, at least according to the questionnaire. The question how to make this attack even more difficult is future work. Another problem is that we have observed that some comments did not have any relevance for the discussion; still they have not been marked as off-topic. By finding ways to reduce the number of or eliminate this kind of comment, the overall quality of the arguments would increase. One way to deal with this problem could be to introduce another category next to ‘repetition’ or ‘off-topic’, namely ‘irrelevant’, and to have a respective new criterion, i.e., participants must not post irrelevant comments. Another solution might be to leave aside arguments without any ratings or follow-up comments when computing proposal scores. This item is a specific example of a larger issue, namely that our model can still be improved. We also see problems with the peer prediction method as an indicator of “honesty”. It has turned to be very complex, in particular when thinking of participants who quickly want to jump into the discussion. A possible solution could be to rely on community consensus instead. But this has not been part of this current study.

As stated before, the evaluation of proposal extensions is a very difficult open issue, considering the diversity of extensions. For instance, we do not see at this point how to decide whether two proposal extensions mutually exclude each other, or could both be implemented. Further, even if we could answer this question, we would have to decide how to select the extensions to be implemented. As mentioned, we have evaluated the extensions by hand in our current study. The fact that nobody from the community has brought up any concerns regarding this could indicate that participants might already be happy with a moderator/elected representative choosing the extensions to be implemented, as long as the proposal with the highest score will be carried out.

Finally, as mentioned, our model is ad-hoc, and improvements are likely to be possible. However, this is not in contradiction to our contributions. In a nutshell, our concern has been to check whether our specific model is useful.

VIII. Conclusions

In this article, we have proposed a novel approach for group-decision making in online settings. It relies on the established principle of deliberation, i.e., collecting and exchanging arguments in order to rank solution options. An essential feature of the approach is that individuals whose behavior is in line with various formal criteria have a higher influence on the decisions. To arrive at a discussion structure that gives way to a ranking of participants and of solution options, we have come up with various extensions of conventional forum structures. We have evaluated our approach by conducting an experiment with a community discussing topics relevant for it. Our overall impression is that the participants have addressed the issues very well. The results we have presented here are from the survey conducted after the experiment. They suggest that that particular community has been satisfied with our forum model and the respective decisions.

References

[1] M. Klein, "Enabling Large-Scale Deliberation Using Attention-Mediation Metrics," Journal of Computer Supported Cooperative Work (CSCW), vol. 21, pp. 449-473, October 2012.

[2] T. Kriplean, J. Morgan, D. Freelon, A. Borning, L. Bennett, "Supporting reflective public thought with considerit", Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work - CSCW '12, 2012.

[3] M. Klein,, "How to Harvest Collective Wisdom on Complex Problems: An Introduction to the MIT Deliberatorium", CCI working paper, 2011.

[4] L. Terveen, and W. Hill "Beyond recommender systems: Helping people help each other". HCI in the New Millennium. J. M. Carroll. New York, Addison-Wesley, 2002.

[5] C. Dellarocas, "The Digitization of Word of Mouth: Promise and Challenges of Online Feedback Mechanisms," Management Science, vol. 49, pp. 1407-1424, 2003.

[6] D. N. Walton, and E. C. W. Krabbe, Commitment in dialogue: Basic concepts of interpersonal reasoning. Albany, NY: State University of New York Press, 1995.

[7] F. H. v. Eemeren, and R. Grootendorst, A Systematic Theory of Argumentation: The Pragma- dialectical Approach. Cambridge University Press, 2003.

[8] S. Chambers, Reasonable Democracy. Ithaca, NY: Cornell University Press, 1996.

[9] J. Cohen, Deliberation and Democratic Legitimacy. The Good Polity: Normative Analysis of the State, eds. A. Hamlin and P. Pettit. Cambridge, UK: Basil Blackwell, 1989.

[10] M. X. Delli Carpini, F. L. Cook, and L. R. Jacobs, "Public Deliberation, Discursive Participation, and Citizen Engagement: A Review of the Empirical Literature," Annual Review of Political Science, vol. 7, pp. 315-44, 2004.

[11] J. Fearson, Deliberation as Discussion. Deliberative Democracy. ed. J. Elster, 44-68. Cambridge, UK: Cambridge University Press, 1998.

[12] J. Fishkin, Democracy and Deliberation. Binghamton, NY: Vail-Ballou Press, 1991.

[13] J. Fishkin, The Voice of the People. Binghamton, NY: Vail-Ballou Press, 1995.

[14] J. Gastil, By Popular Demand. Berkeley: University of California Press, 2000.

[15] A. Gutmann, and D. Thompson, Democracy and Disagreement. Cambridge, MA: Harvard University Press, 1996.

[16] S. Isenmann, and W. Reuter, "IBIS - a Convincing Concept . . . But a Lousy Instrument?", Conference on Designing interactive systems processes, practices, methods, and techniques, 1997.

[17] S. Shum, "Cohere: Towards web 2.0 argumentation," International Conference on Computational Models of Argument, 2008.

[18] D.N. Walton, and C.A. Reed, "Diagramming, Argumentation Schemes and Critical Questions", 5th International Conference on Argumentation (ISSA'2002), SicSat, Amsterdam, pp. 881-885, 2002.

[19] C. Lampe, and P. Resnick, "Slash(dot) and burn: distributed moderation in a large online conversation space," Conference on Human Factors in Computing Systems (CHI), Vienna, Austria, ACM Press, pp. 543-550, 2004.

[20] D. Schuler, "Online civic deliberation with E-liberate, in Online Deliberation: Design, Research, & Practice", edited by Davies, Center for the Study of Language and Information (CLSI), Stanford, California, pp. 293-303, November 2009.

[21] S. Park, S. Kang, S. Chung, S. Song, Junehwa, "NewsCube : Delivering Multiple Aspects of News to Mitigate Media Bias," Conference on Human Factors in Computing Systems, 2009, pp. 443-452, 2009.

[22] T.Kriplean, J. Morgan, "REFLECT: Supporting Active Listening and Grounding on the Web through Restatement," Conference on Computer Supported Cooperative Work, 2011.

[23] S.Faridani, E. Bitton, K. Ryokai, K. Goldberg, "Opinion space: a scalable tool for browsing online comments," SIGSHI Conference on Human Factors in Computer Systems, pp. 1175-1184, 2010.

[24] S. Tanasijevic, K. Böhm, Technical Report, Karlsruhe Reports in Informatics, 2013., http://digbib.ubka.uni-karlsruhe.de/volltexte/1000034625

[25] N. Miller, P. Resnick, R. Zeckhauser, "Eliciting Informative Feedback: The Peer-Prediction Method," Management Science, vol. 5, 2005.

[26] R. Jurca, B. Faltings, "Minimum payments that reward honest reputation feedback," ACM conference on Electronic commerce, 2006.

[27] https://www.phpbb.com/

[28] http://shakuras.ipd.uni-karlsruhe.de/dbsforum/

[29] http://shakuras.ipd.uni-karlsruhe.de/dbsforum/pdf/Questionnaire.pdf