Behavioral Strategies in Online Forums with Different Feedback Types

Sanja Tanasijevic
Karlsruhe Institute of Technology
Karlsruhe, Germany
Email: sanja.tanasijevic@kit.edu

Klemens Böhm
Karlsruhe Institute of Technology
Karlsruhe, Germany
Email: klemens.boehm@kit.edu

Karl-Martin Ehrhart
Karlsruhe Institute of Technology
Karlsruhe, Germany
Email: ehrhart@kit.edu

***Abstract*—Forums for deliberation, i.e., coming to decisions and solutions for community problems, play an important role in public life. Methods for evaluating the content of such forums are becoming more and more significant. A promising direction is to introduce feedback options of different types , i.e., participants can assess a post by someone else according to different criteria, such as agreement/disagreement, originality or relevance for the topic currently discussed. However, participants in such forums may have specific interests such as earning themselves a high weight or earning high scores for 'their' arguments. Feedback options of different types now give rise to new kinds of strategic behavior, to back up one's specific interests and to push one's opinion. For instance, a participant could label an argument that he does not like as 'not original', in order to bog it down. To study this kind of behavior, we have built a respective game-theoretic model. The model incorporates an evaluation scheme, as follows: (1) users are assigned weights based on criteria such as originality of their posts according to feedback by others; (2) posts are scored based on the rate of agreement/disagreement feedback they have obtained and the weights of raters and authors. The model lets us study the following questions: When exactly does untruthful rating behavior pay off, and is truthful behavior an equilibrium strategy? A core insight is that the strategy 'rate posts always truthfully' is an equilibrium strategy. Further, our evaluation scheme is robust towards untruthful behavior of participants in many cases.**

I. Introduction

Forums for deliberation, i.e., coming to decisions and solutions for community problems, are important to deal with issues of common interests. They allow collecting the opinion of communities regarding any issue, on the level of a state, a municipality, or a group. However, the problem of evaluating content prevails. It would help a lot to single out comments[1] which are not related to the topic of the discussion or repetitions of previous comments. Other comments 'of interest' are those that are not understandable or that are offensive or provocative. At the same time, and orthogonally to these dimensions, it is important to know whether a community agrees or disagrees with a certain comment. Note that one-dimensional feedback options such as 'thumbs up', 'like' buttons etc. are too undifferentiated to this end. If someone clicked, say, a 'thumbs down' button in a deliberation forum, it would be unclear whether he disagreed with the argument or had an objection on the formal level, e.g., the argument is a repetition of a previous one. In many other contexts in turn, e.g., entertainment, online documentation pages, this difference does not exist, and one-dimensional feedback is conclusive. A viable solution to that problem in deliberation forums seems to be feedback where different types are feasible in turn, i.e., participants can assess a post by someone else according to different criteria.

Discussants have different interests, which motivate them to issue feedback on posts. In the presence of feedback of different types, they can behave strategically to back up their specific interests and to push their opinion, cf. [1].

Example 1: Consider a forum discussion on municipality budget savings such as http://essen-kriegt-die-kurve.de. One of the proposals discussed there has been to raise the tax for pet owners. Further, think of a new, solid argument in favor of this tax increase. A forum participant who is a pet owner can now behave in the following ways: (a) Act truthfully and give feedback honestly, considering the argumentation. (b) Issue feedback strategically, e.g., mark the post as a repetition of a previous one, in order to bog it down, to protect his own interest or to push his opinion. Obviously, the opposite case exists as well, namely a post that is a repetition of a previous one, but may not be marked as such by its supporters.

In this article, we study this specific setting, namely *various rating strategies* in the presence of *feedback of different types*. The core research questions are when exactly untruthful rating behavior may pay off, and how this can be avoided. At first sight, it might seem worthwhile to examine whether truthful behavior is an equilibrium strategy. However, there are some situations where untruthful behavior obviously is beneficial.

Example 2: Think of a state of the discussion forum with two Comments A and B. No participant has rated them yet. A is a repetition of a previous comment, whereas B is not. Participant P agrees with A, but disagrees with B. He could 'overlook' that A is a repetition and rate it with "Agree", and he could bog down Comment B by rating it as "Repetition". With any reasonable scoring scheme for comments, P gains an advantage from behaving strategically.

This example has several important implications: P is able to gain an advantage because the feedback by others (or the fact that not one else has given feedback yet) is known to him. Consequently, we focus on forum designs where less or no information on the feedback the various comments have received so far is revealed. Specifically, we study a setting where ratings are not published at all until the time window, when issuing feedback is possible, ends. Such a setting also is

[1]In this article, comment, argument and post are used as synonyms. We also use feedback and rating synonymously.

appropriate when the objective indeed is to collect the true opinions of participants and to avoid herding behavior. As participants do not see feedback by others, certain system states are indistinguishable for them. In consequence, studying whether truthful behavior is an equilibrium strategy is more meaningful, i.e., is the *expected* utility maximal when behaving truthfully, assuming that everybody else acts truthfully. A subsequent question is how robust this equilibrium is in various respects, including the rate of comments which are misperceived, e.g., perceived as a new post even though it is a repetition, or the share of trembling, i.e., participants giving untruthful feedback, be it by mistake, be it on purpose. *A trembling hand equilibrium* is one that takes the possibility of off-the-equilibrium play into account by assuming that players, through a "slip of the hand" or **tremble**, may choose unintended strategies, albeit with low probability [2].

A. Challenges

The problem investigated here is challenging, for several reasons: First, we have to build a formal model of forum discussions. Models of social systems are complex, with imprecise, incomplete and inconsistent theories [3]. Our model should be sufficiently exhaustive and representative to allow for meaningful conclusions, it must not be overly complex. The objective is to mimic the discussion structure, e.g., who has authored a comment or has posted a feedback item, and of which value. In particular, the model should feature the dynamics of forum discussions and reflect the possible actions, moves of participants and probabilities of these moves.

A specific question is how utility should be defined in this current context. In our previous work where the evaluation had consisted of user experiments [4], we have followed a two-step weighting and scoring scheme. That is, in a first step, participants have been weighted according to various criteria, such as number of comments they have posted. If indicators such as this number are high, the weight of the respective participant tends to be high as well, to express appreciation for, in this example, more activity. In a second step, comments are scored, not only taking the feedback given by participants into account, but also the weights of the respective participants. Thus, the rationale behind the weights is to drive participant behavior in directions that are desirable from the perspective of a forum organizer, by giving 'good' participants a higher influence on the comment scoring. As mentioned earlier, participants now may have different motivations to participate, including (a) pushing their perspective on things, or (b) distinguishing themselves as good members of the community, while not really being interested in the discussion outcome. This translates to different notions of utility: (a) implies that arguments one is supportive of having high scores yields high utility. In contrast, high utility in the case of (b) goes along with a high weight. An issue to be observed not only is to model these different variants of utility appropriately; it also includes checking whether truthful behavior constitutes an equilibrium in these different cases.

On a technical level, as each participant has several options to give feedback on a comment, the number of system states already is intimidating for few comments and medium-sized communities. Not all states can be inspected explicitly, as would be necessary in a conventional game-theoretic analysis. So the state space needs to be narrowed down by much, to give way to the analysis envisioned. In principle, one way to do this, in line with the fact that we are seeking an equilibrium, is sampling the set of states. However, ensuring that such a sampling is not biased is not trivial.

B. Contributions

Our contributions are as follows. We show, for one meaningful class of weighting and scoring schemes, that rating posts truthfully constitutes a symmetric equilibrium (in the game-theoretic sense). We show this for different definitions of utility, namely one based on a ranking of comments, and one based on a ranking of participants. The model we have built for this purpose is sufficiently general and includes the case that posts can be misunderstood or overlooked. On the technical level, our approach is the following one: While the number of states is daunting, we observe that they can be grouped into relatively few equivalence classes. Two states are equivalent from the perspective of a discussant if he cannot distinguish them. In our setting, this is because feedback by others is not published right away. For each equivalence class, we compute the expected utility of truthful behavior as well as of other strategies for a sample of states from that class that is statistically significant. For settings where the expected utility of truthful behavior is higher than the one of other behavior for all equivalence classes, we conclude that there is an equilibrium. Our analysis reveals that this is the case in all relevant settings. Next, we show that this equilibrium is robust against a negligible amount of trembling. We quantify the extent of trembling that can be tolerated, by varying the relevant parameters systematically.

To our knowledge, this article is the first to examine the effects of strategic behavior in forum discussions in the presence of feedback of different types. It is an initial stab at the problem and is obviously incomplete. Our various omissions include the following ones: (a) We compare truthful behavior only with a small number of alternative strategies, including 'always untruthful'. While we do not expect any difficulties with our model/implementation when using mixed strategies ('sometimes truthful-sometimes untruthful') as a reference point instead, other strategies that, for instance, take the content of posts into account are not covered here. (b) We limit the study to one class of weighting/scoring schemes, the two-step approach mentioned earlier. Regarding this point however, we do not foresee any complications when extending the study to other scoring schemes; this is because our approach/implementation encapsulates the scoring scheme well. (c) We only consider one kind of unwanted behavior: We focus on repetitions of arguments that have already been posted. The unwanted behavior examined here is using the respective feedback strategically, cf. Example 1. Studying other variants of unwanted behavior instead, such as using feedback for posts that are offensive or are off-topic strategically, should be more or less identical to what we have done, except that possible actions of participants need to be modeled differently. On the other hand, we have not examined other kinds of unwanted behavior, e.g., *posting* repetitions strategically, let alone the effects of different kinds of unwanted behavior being feasible at the same time. Finally, the method used here is *bounded* model checking. While this is a method that is generally accepted in computer science, this means that the

results we have obtained only hold for models of a certain size (relatively few participants and few posts) with certainty.

Paper outline: We describe our formal model, as well as the weighting and scoring scheme in Section III. We introduce different strategies that are conceivable in the context of repetitions. Next, we present the evaluation setup and the utility functions used to assess the benefit of the strategies studied. In Sections V and VI, we present our results and conclude.

II. Related Work

Work directly related to the strategic behavior of individuals in settings with feedback of different types is rare. This section takes a somewhat broader view on related work, focusing on work that examines participant behavior in online portals using game theory. With all contributions mentioned in the following, the problem studied is different from ours (no feedback of different types); we do not mention this point in the remainder of this section any more.

The work presented in [5] is one of the first proposals of a game-theoretic model for deliberation. The author describes a dynamic model, for a non-cooperative game. Another article by the same author [6] offers insights regarding the dynamics of the deliberation process and its steady point, the so-called deliberation equilibrium. [5][6][7] use the concept of deliberation based on expected utility.

Important issues in research are the motivation to contribute and optimal designs of reward systems in online social systems, e.g., forums or knowledge-sharing websites, and there are many recent innovations regarding such platforms. [8] proposes a game-theoretic framework to study the dynamics of a social media network where contribution costs are individual, but gains are common, and users are rational selfish agents. In this project, incentives are explicitly quantifiable (monetary or virtual credit). In our context in turn, the incentives for taking different strategies are more involved (ranks of the participant or scores of the arguments), and calculating the expected utility of participants is more complex. [1] proposes a ranking mechanism that maximizes the utility of the 'game owner' and incentivizes participants to give high-quality contributions. This article highlights the important point that game owners and participants have different interests: The game owner wishes to optimize an objective, typically a function of the number and quality of contributions received. Potential contributors in turn think strategically, i.e., decide whether to contribute or not to selfishly maximize their own utility, e.g., visibility in their respective communities. The authors present a game-theoretic model to study whether contests aiming at the best contributions give way to optimal outcomes. Being aware of this distinction, we target at valuation schemes that nudge discussants to contribute and give original arguments and honest feedback. [9] describes a game-theoretic model of a crowdsourcing contest and studies how to split a prize budget among contestants to achieve the so-called "maximum equilibrium effort".

[10] examines behavior of users on a Chinese web-based knowledge-sharing market, Taskcn.com. They find a significant variation in the expertise and productivity of the participating users: A very small core of successful users contributes nearly 20% of the winning solutions on the site. A user who is successful not only manages to win multiple tasks, but also to increase his win-to-submission ratio over time. This is in line with our underlying assumption that participants do behave strategically; in particular, they pick tasks whose expected level of competition is lower. [11] provides a game-theoretic model of multiple simultaneous crowdsourcing contests where agents select among, and subsequently compete in, several contests offering various rewards. Authors model crowdsourcing contests as all-pay auctions with incomplete information on contestant skills.

While all these projects rely on game theory to represent a specific kind of social system, our focus is unique in that we study forum discussions with multiple feedback options and strategic opportunities arising from.

III. Formal Model

Game theory is commonly used to analyze settings where different decision makers meet and might have conflicts. Dynamic models of deliberation embed the theory of non-cooperative games. [5] Forum participants are players. Nature (i.e., certain probabilities that are exogenous parameters) determines whether a comment that the author has intended to be a new argument is perceived as such or as a repeated argument; same with comments that are intended as repetitions. Nodes/intermediate states describe the progress of the game/discussion with arguments and ratings generated. The act of generating an argument or a rating is a move, an option available to a player at a certain state of the game. Participants choose strategies based on their expected utilities. A utility function quantifies the benefit or loss of participants. In our case, a participant might gain a benefit from lowering the scores of comments he disagrees with, to give an example.

The formal model envisioned and proposed in what follows features the following notions: different types of feedback; a weighting scheme, to assess participants; a scoring scheme, to evaluate comments. The model features posting comments or ratings of comments authored by others. A participant can post a rating of the following values: Agree, Disagree, Repetition. A participant does not have to rate a certain comment.

To illustrate the nature of the game, we use Figure 1. A participant can post a comment intended to be an original one ('New') or a repetition of a previous one ('Repetition'), denoted by the dotted box labeled 'Intention'. The intention of the author is not his decision between moves, but rather a nature of the game described by the probability of repetitions, p_r (exogenous parameter). Thus, the intention is determined by a so-called *random move of nature*. We do not (and do not need to) concern ourselves with the true nature of a comment, only the intention of the author is relevant. Rectangles represent participants, whereas eclipses stand for comment states. Raters can perceive comments as intended by authors or misperceive them again by the nature of the game, see the states within the dotted box labeled 'Perception'. p_m is an exogenous parameter referring to the probability of misperception. For instance, if p_m is set to 0.5, every other comment is misperceived, compared to its nature intended by the author.

Regarding the rating behavior of participants, the following points are important. A participant can post at most one rating per comment, only for comments authored by other

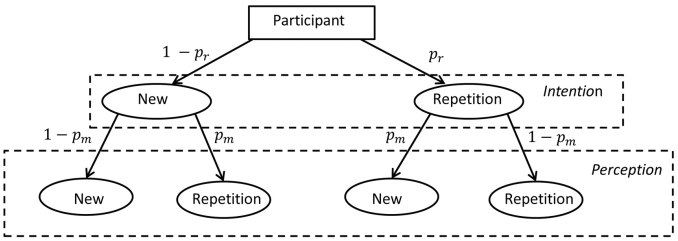

Fig. 1. Nature of the game

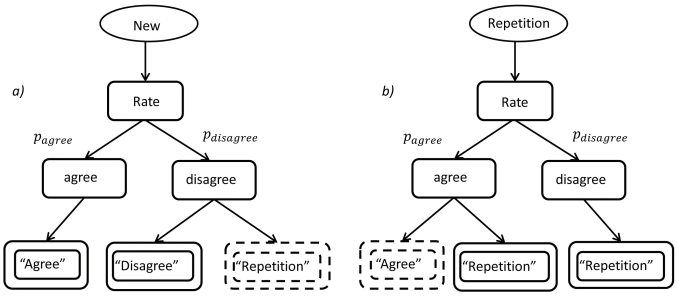

Fig. 2. Participants moves for new/repetition comments

participants. When doing so, he can follow different strategies, e.g., 'always truthful' or 'always untruthful', as we discuss later.

Figure 2 a) shows the possible moves of a participant when perceiving a comment as new, Figure 2 b) shows the same for a comment perceived as repetition. 'Rate' means that the participant has rated the current comment. While 'Agree' and 'Disagree' (without double quotation marks) refer to the true perception of the rater, "Agree", "Disagree", "Repetition" are the ratings actually posted. p_{agree}, $p_{disagree}$ are the probabilities (exogenous parameters) that a participant agrees or disagrees with a comment. Looking at the left graph, when a rater agrees with a new comment, he obviously maximizes his benefit by rating it truthfully, i.e., issuing "Agree". So any alternatives are not represented explicitly. On the other hand, when disagreeing with a new comment, a rater can behave truthfully and issue "Disagree". When behaving untruthfully, he issues "Repetition". His rationale would be to prevent the comment from being taken as a valuable argument. The other edge labels can be ignored for the time being; we will cover them later. Similarly, the right graph illustrates moves available to participants when perceiving a comment as a repetition. In both figures, moves representing the *untruthful strategy* are highlighted with long dashed line. Observe that Figure 2 represents possible moves of participants. If they always behave truthfully or always untruthfully, they decide for *pure strategies*. However, participants may deviate from their pure strategies, and this is referred to as *mixed strategy* (i.e., a certain probability distribution over the set of pure strategies). Note that participants cannot see other ratings (or comments scores, which we will discuss in detail later). The reason behind this design decision has been to prevent participants from influencing each other (herd behavior) and to ensure that comment scores are realistic indicators of the community opinion.

A. Weighting Scheme

The next constituent of our model is a scoring scheme, i.e., a function that assigns values to posts, and that takes feedback issued by participants into account. A broad range of such functions is conceivable and could in principle be used here. In what follows, we describe the specific one that we have explicitly tested, see Section IV. However, we stress that any other scoring function can be evaluated without difficulty. The reason is that it is encapsulated well both regarding our approach and its implementation.

In this study, we distinguish between weights and scores, as follows: While both are supposed to facilitate an evaluation, weights are characteristics of participants, based on criteria such as rate of repetitions posted according to feedback by others. Scores in turn are characteristics of comments, based on the degree of agreement in the community as well as on the weights of authors and raters. Weights are part of our approach to enforce certain kinds of desired behavior and discourage participants from unwanted behavior such as posting repetitions. In other words, by behaving in an assimilated manner, participants can gain a higher influence on the comment scores.

In this work, we assume that there is only one kind of unwanted behavior, namely repeating arguments which have already been posted. We also confine our study to 'rating behavior', i.e., generating comments (and behaving strategically when doing so) is not considered either. In consequence, in contrast to our previous experimental work where we had many so-called indicators influencing the weights, to cover other kinds of unwanted behavior, we only look at the following criteria in what follows:

- **Originality**. If a participant authors comments which are not (or only rarely) rated as repetitions, the originality indicator will have a high value.

- **Rating consensus**. This criterion quantifies the degree of consensus with other raters on whether a comment is a repetition or not. If the rates of a participant frequently are in line with most rates whether a comment is a repetition, this indicator is high. The rationale is to assess whether participants generate ratings truthfully. This criterion only takes the "Repetition" rating into account, but not "Agree" or "Disagree". The reason is that participants should be free to post their true opinion regarding the content of arguments, irrespective of what the majority thinks. With repetitions in turn, we hypothesize that there is some objective truth, which participant ratings should meet.

We refer to the indicator for originality as $orig(j)$. $R^{subject}(j)$ is the set of ratings for comments authored by Participant j. $R^{subject}_{repetition}(j)$ is the set of "Repetition" ratings for comments authored by Participant j.

$$orig(j) = \begin{cases} \text{null} & \text{if there are no posts or no ratings,} \\ 1 - \frac{|R^{subject}_{repetition}(j)|}{|R^{subject}(j)|} & \text{otherwise.} \end{cases} \quad (1)$$

We refer to the consensus indicator as $consensus\_ratings(j)$. Each rating r, repetition or opinion

rating (agreement/disagreement) posted by Participant j, is evaluated based on its degree of consensus in the set of ratings that comment has received. Finally, the average is calculated for all ratings of Participant j. $aname$ is a function that returns one of the following values: Repetition, Agree/Disagree. These values form the set $range(aname)$. Thus, $r.aname$ refers to particular rating posted by a rater for a comment. $r.comment$ is the comment the rating refers to. The set of ratings issued for the comment of a particular value such as "Repetition" or "Agree"/"Disagree" is denoted by $R^{subject}_{aname}(r.comment)$. $R^{create}_{repetition}(j)$ is the set of all "Repetition" ratings Participant j has issued. $R^{subject}(r.comment)$ is the set of all ratings of the comment. First, we calculate a value for each rating issued by Participant j.

$$share(r, aname) = \frac{\left|R^{subject}_{aname}(r.comment)\right|}{\left|\sum_{a \in range(aname)} R^{subject}_{aname:a}(r.comment)\right|}$$

Then we calculate averages of the scores of all ratings' per value, i.e. "Repetition" or "Agree"/"Disagree".

$$consensus\_ratings(j, aname) =$$

$$\begin{cases} avg_{r \in R^{create}_{aname}(j)}\left(share(r, aname)\right) & \text{if } R^{create}_{aname}(j) \neq \oslash, \\ 0 & \text{otherwise.} \end{cases} \tag{2}$$

Finally, the value of this indicator for Participant j is the average of all rating values:

$$consensus\_ratings(j) =$$

$$avg_{a \in range(aname)}\left(consensus\_ratings(j, a)\right)$$

Having formalized the various criteria that are relevant here, the remaining step is to compute participant weights based on his indicator values. Several aggregation functions are conceivable; the one used here is the minimum function.

$$weight(j) = \min(orig(j), consensus\_ratings(j))$$

Having sorted participants by weight, the resulting list features the weight rank of each participant: The higher the participant appearance in this list, the higher his weight rank.

B. Scoring Scheme

A scoring scheme evaluates comments based on the ratings received and, in our case, the weights of raters and authors. Recall that the rationale behind feedback of different types is quality control, i.e., to 'sort out' certain unwanted comments. In our case, these are repetitions of previous posts. We propose to ignore comment having received more than 50% "Repetition" ratings, and no score is computed for them. Otherwise, the score of a Comment k is as follows:

$$score(k) = \nu(k) \cdot \left(\frac{weight(author(k)) + \sum_{r \in R^{subject}_{opinionagree}(k)} weight(issuer(r))}{weight(author(k)) + \sum_{r \in R^{subject}_{opinion}(k)} weight(issuer(r))} - 0.5 \right)$$

where

$$\nu(k) = \frac{weight(author(k)) + \sum_{r \in R^{subject}_{opinion}(k)} weight(issuer(r))}{\max_{k' \in K}(weight(author(k')) + \sum_{r \in R^{subject}_{opinion}(k')} weight(issuer(r)))}$$

$weight(author(j))$ is the weight of the author of Comment k. $weight(issuer(r))$ is the weight of a rater of Comment k, the one who has generated Rating r. $R^{subject}_{opinion}(k)$ is the set of all agreement and disagreement ratings for Comment k. $R^{subject}_{opinion:agree}(k)$ is the set of all agreement ratings for Comment k. The first factor in the equation sets the degree of agreement to a value in the range $[-0.5, 0.5]$. For instance, if all ratings of a comment are "Agree", the score will be 0.5, which is maximal. Additionally, scores are normalized with $\nu(k)$. This is the ratio of the weights of the author and the raters of k over the maximal sum of author and raters weights of all comments. In other words, if a comment receives a few positive ratings, its score should be comparable to the one of comments which received a lot of attention from the community, but not necessarily in the form of positive feedback throughout.

IV. Model Evaluation

The evaluation of the formal model is done in a comprehensive way, with an emphasis on the following points. We investigate if the 'always truthful' strategy is an equilibrium strategy. Furthermore, we want to assess the robustness of the proposed weighting and scoring scheme with different rating strategies. Namely, we want to check if the always untruthful strategy pays off. Another issue is that the number of possible outcomes of the game investigated here is huge (and even infinite if the number of participants or posts is not bounded); we cannot explicitly inspect each of them.

A certain strategy is a (symmetric) equilibrium strategy if it yields maximal utility for a player/participant, provided that all other participants follow this strategy. A strategy is an equilibrium strategy if the *expected utility* of a player is maximal under that provision. Thus, in a nutshell, to check whether a certain strategy ('always truthful' in our case) yields an equilibrium, we assume that all participants follow it, except for one participant, the so-called *controlled participant*. We refer to the other participants as *synthetic participants*. In order to check if "always truthful" constitutes a symmetric equilibrium (i.e., all players play the same strategy), we have to check whether the strategy "always truthful" is the controlled participant's best response (in the sense of maximum expected utility) if all other players (i.e., the synthetic participants) play "always truthful". We then generate a system state that is

'almost complete'. This means that the actions of all synthetic participants, i.e., which comments they have given feedback to, and which values, are specified, except for the controlled participant. We then generate two completions of this state: One includes the actions of the controlled participant following the 'always truthful' strategy, the other one includes his actions when following the reference strategy, e.g., 'always untruthful'. We compute the utility of the controlled participant in these two cases. We refer to such a sequence of steps as simulation run. If the utility of 'always truthful' is higher (not lower) than the other one, this is a good sign. We repeat these simulation runs – which are random processes, as we will explain right away, i.e., the new state will most likely be different from the previous one – and again compare the utilities. We keep doing this until there is some statistical significance that the controlled participant can expect a higher utility from one of the strategies.

In more detail, as feedback by others is hidden from participants, they cannot distinguish between certain states. This gives way to a definition of equivalence of states. We aim to show that, in each equivalence class, the expected utility of truthful behavior is higher that of other strategies. We do this by repeating the procedure outlined in the previous paragraph for each class; we declare success only if 'always truthful' yields an expected higher utility for all classes. Next, if this is successful, we want to quantify the robustness of this equilibrium in various respects, including trembling, i.e., participants do deviate from 'always truthful' to some extent.

A simulation run is governed by the following parameters:

(1) Number of participants ($num\_participants$), comments generated ($num\_comments$) and maximum number of ratings generated ($num\_ratings$).

(2) Probability of misperceiving a comment (p_m), i.e., rater perception of a comment differs from the intention of the author (see Figure 1). Perceiving a new comment as a repetition and repetition as new can happen with certain probabilities. To keep the setting simple, we use one value for these probabilities. In a first investigation, we set this value to zero. We then vary the parameter, i.e., add noise to the perception of raters, to make the model more realistic.

(3) Mixed strategy of synthetic participants, i.e., whether a synthetic participant rates a comment that he perceives as new not as a repetition and vice versa (see Figure 2). Note that this is only relevant when studying trembling. Namely, we are interested in quantifying the extent of trembling the scoring scheme can tolerate.

(4) Probabilities that determine the rating behavior of participants, i.e., agreement/disagreement ratio in the community (p_{agree}, $p_{disagree}$). By varying this ratio, we mimic more or less homogenous forums.

Considering the stochastic nature of the approach, due to the sampling of the possible states we can also vary other settings. To study the robustness of the weighting scheme, we can also examine other weighting functions. Finally, to stress test the equilibrium, we observe trembling in the behavior of the controlled participant.

A. Implementation of the Formal Model

The formal model we have proposed has the following elements: participants, comments, ratings and equivalence classes. As part of our formalization, we introduce the following notation. K is the set of comments. P is the set of participants, $p\_controlled$ is the controlled participant. Next, we want to distinguish between comments the controlled participant agrees with and disagrees with, respectively; the so-called valuation function $v : K \to \{+, -\}$ accomplishes this. Note that this is different from the feedback actually given by the controlled participant, v reflects his true opinion. (Notation for the feedback actually given will follow.) K^+, K^- are the sets of comments the controlled participant agrees and disagrees with, respectively. Next, the partial function pf, referred to as perception function, states whether the controlled participant perceives a comment as new or as a repetition of a previous comment; $pf : K \to \{new, repetition\}$. pf and v are orthogonal to each other; any combination of values regarding a comment is possible.

Ratings ("Repetition", "Agree", "Disagree") are randomly granted to comments, following Figure 2. To actually generate ratings we use probabilities of misperception (p_m), agreement and disagreement (p_{agree}, $p_{disagree}$), all specified a priori.

We differentiate between *complete states regarding a set of comments* and *intermediate states*. A complete state is one where all participants have given their feedback regarding the comments (or they have chosen not to give feedback regarding some comments), except for the respective authors. An intermediate state is one where the information whether feedback is given, and how this feedback looks like, is missing for some (comment–participant) combination. We refer to an intermediate state where only the information from the controlled participant is missing as *almost complete*. Finally, a state also includes a valuation function and a perception function, though we refrain from explicitly representing these functions at times, to avoid clutter in the presentation.

In what follows, we use a tabular representation of almost complete states (leaving aside the valuation function and the perception function), referred to as action matrix: Each column corresponds to a comment, each row to a (synthetic) participant. Each cell contains the move of the participant regarding the comment. The following moves are available: (1) write a comment (w); (2) rate a comment with ratings "Agree" (a), "Disagree" (d), "Repetition" (r); and (3) none.

So we can represent each cell as a vector with four Boolean components where value 1 occurs exactly once. Further, value 1 must occur exactly once per column at first place (w), i.e., there is exactly one author per comment.

$$\begin{bmatrix} (1,\,0,\,0,\,0) & (0,\,1,\,0,\,0) & (1,\,0,\,0,\,0) \\ (0,\,0,\,1,\,0) & (1,\,0,\,0,\,0) & (0,\,0,\,0,\,1) \end{bmatrix} \tag{3}$$

Figure 4. Example of an action matrix

The action matrix in Figure 4 for two participants $(P1, P2)$ and three comments $(K1, K2, K3)$ serves as an illustration. Here, Participant $P1$ has posted comments $K1, K3$ and has generated an "Agree" rating for $K3$, whereas participant $P2$ has posted $K2$ and "Disagree" and "Repetition" ratings for

$K1$ and $K3$, respectively. Obviously, $K2$ is the only comment that has not received any rating yet.

B. Formalization

To continue the formalization, we introduce the following notation. A is the action matrix. It corresponds to a partial function state that is defined as *state*: $K \times P \backslash \{p\_controlled\} \rightarrow \{w, a, d, r\}$. The functions *untruthful*: $K \times P \rightarrow \{w, a, d, r\}$ and *truthful* of the same type are extensions of state: They also include the ratings by $p\_controlled$, according to the 'always untruthful' and 'always truthful' strategies, respectively. $score(k, untruthful)$ is the score of Comment k computed with the values returned by function *untruthful*. The notion just introduced will help us to formalize the notion of utility. A utility function is a function that has a valuation function v and an action matrix extended with the ratings of the controlled participant following as input and returns a utility value; we refer to these values as $utility(v, untruthful)$ and $utility(v, truthful)$ for those two strategies. We will describe instantiations of *utility* later.

We have already introduced equivalence classes informally at an abstract level; we now provide more details. Equivalence classes comprise states the controlled participant cannot distinguish from each other: By definition, two states are equivalent if the comments are identical[2], as are the valuation functions and the perception functions. To illustrate, think of a very simple setting consisting of only one Comment k. In this case, there are four equivalence classes:

(1) The controlled participant perceives Comment k as a repetition and disagrees with it.

(2) He perceives k as a repetition and agrees with it.

(3) He perceives k as new and disagrees with it.

(4) He perceives k as new and agrees with it.

We can represent an equivalence class of states as a vector ($z1$, $z2$, $z3$, $z4$) where:

$$z1 = |k \in K : v(k = + \wedge pf(k) = new)|$$

$$z2 = |k \in K : v(k = - \wedge pf(k) = new)|$$

$$z3 = |k \in K : v(k = + \wedge pf(k) = repeat)|$$

$$z4 = |k \in K : v(k = - \wedge pf(k) = repeat)|$$

To illustrate, with 10 comments the number of possible equivalence classes is 286. Namely, there are four different classes assigned to 10 comments. The total number of classes is calculated as the number of combinations with repetitions[3].

[2]Note that we ignore the content of comments, so this requirement means that there must be the same number of comments in both cases.

[3]http://www.statlect.com/subon2/comcom1.htm

C. Utility functions

The next step is calculating the utility of a participant, the controlled participant in our case, for a given state, for different strategies such as 'always truthful' or 'always untruthful'. Different utility functions are conceivable. To illustrate, a possible benefit of a participant could be pushing his opinion, i.e., scores of comments he agrees with end up to be higher with a certain strategy. At the same time, the participant might not be interested at all in his weight. The opposite perspective, i.e., participants do not care about comment scores at all, but strive for high weights, is possible as well, in particular in settings where participant weights are displayed prominently for the entire community. In our evaluation, we will study four different utility functions. More specifically, in some cases, instead of having one explicitly defined utility function that quantifies the usefulness of a strategy, we have found it more convenient to define a *relative utility function* that takes two strategies as input and yields a positive result if the first strategy is better than the second one (Items (1) and (3) in what follows). In the following, we let s denote the strategy under observation and s' the reference strategy. The following list is an overview, followed by a formalization of each of them:

(1) We compare comment scores for the strategy under observation (s) and the reference strategy (s'). If there are many comments in K^+ with a higher score for s than for s', the relative utility, referred to as $utility\_count(v, s, s')$, is high. Note that the order of the parameters plays a role – the strategy under observation is listed first. Similarly, if there are many comments in K^- which have lower scores for s than for s', it will be high as well.

(2) Sum up the scores of comments in K^+; do the same for K^-. If the difference of these two values is high, then the utility $utility\_sum$ is high.

(3) We now quantify how well the different strategies help to resolve repetitions. More specifically, if there are comments in K^+ not resolved as repetition when the controlled participant uses s, but identified as such in the other case, the value of $utility\_rep(v, s, s')$ is high. Analogously, if there are comments in K^- which are resolved as repetitions with s, but not with s', the value is high as well.

(4) $utility\_rank$ is the weight rank of the participant when following a certain strategy.

(1) The relative utility function $utility\_count$ has the valuation function v, the observed strategy s and the reference strategy s' as arguments.

$$\begin{aligned} utility\_count(v, s, s') = \\ \big|\{k \in K^+ : score(k, s) \geq score(k, s')\}\big| + \\ \big|\{k \in K^- : score(k, s) \leq score(k, s')\}\big| \end{aligned} \quad (4)$$

(2) The utility function $utility\_sum$ quantifies the benefit of the controlled participant based on the scores of comments

he agrees and disagrees with.

$$utility\_count(v,s) = \sum_{k \in \{k \in K^+ : score(k,s) \neq null\}} score(k,s) - \sum_{k \in \{k \in K^- : score(k,s) \neq null\}} score(k,s) \quad (5)$$

(3) The relative utility function $utility\_rep$ has the valuation function v, the observed strategy and the reference strategy as arguments.

$$utility\_rep(v,s,s') = \left|k \in K^+ \wedge score(k,s) \neq null \wedge score(k,s') = null\right| + \left|k \in K^- \wedge score(k,s) = null \wedge score(k,s') \neq null\right| \quad (6)$$

(4) This utility function takes the weight of the controlled participant as success criterion. $weight(p\_controlled, s)$ quantifies the controlled participant's benefit when using strategy s.

$$utility\_rank(v,s) = weight(p\_controlled, s)$$

D. Simulations

We initialize the formal model with $num\_participants$ number of participants and $num\_comments$ number of comments completely randomly authored by the participants. Then we generate equivalence classes. We generate a certain number of ratings, e.g., we set the maximum number of ratings, and the number of ratings generated depending on the number of comments and raters. I.e., we must obey certain rules, such as that a rater can post one rating per comment posted by other participants. The ratings are completely randomly assigned to participants as raters.

In this section, we provide further details on the simulations conducted and the respective setup. Almost complete states are randomly generated, and the utilities of two strategies, 'always truthful' and 'always truthful', are compared. If 'always truthful' yields a higher utility, a respective counter is increased, otherwise another counter. This is repeated until a certain statistical significance is reached. Note that we do observe only pure strategies, 'always truthful' and 'always untruthful', thus, *untruthfulness of generated ratings* is set to 0. This is repeated for each equivalence class. We now say how we have computed the number of simulation runs necessary per equivalence class. Each simulation run can be seen as a Bernoulli trial, where p is the probability that 'always truthful' pays off. So the hypothesis that we want to reject, with a certain level of confidence, is $p \leq 0.5$. We replace this hypothesis with the one that $p = 0.5$. Namely, if we can reject this hypothesis with a certain level of confidence, given that 'always truthful' pays off in more simulation runs than the reference strategy, then it is clear that smaller values of p are even less likely.

So let $X \sim B(n, 0.5)$. Recall that $Prob(X \leq t) = \sum_{i=0}^{|t|} \binom{n}{i} \cdot p \cdot (1-p)^{n-i}$. In this formula, n is the number of simulation runs carried out so far for the current equivalence class, and t the number of these runs where 'always truthful' has been better than the reference strategy. The probability returned by the formula must be high to reject the hypothesis, e.g., at least 90 percent. Thus, after a certain number of simulation runs for a class, we compute that probability for the current values of n and $num\_comments$. Once that probability threshold is reached, we can stop examining the current class. Table 1 is a summary of the default parameters of our setup.

Number of participants ($num\_participants$)	4
Number of comments ($num\_comments$)	10
Maximum number of ratings ($num\_ratings$)	30
Misperception (%)	0
Comments without ratings	0
Probability of agreement/disagreement (%)	60/40
Untruthfulness of generated ratings	0
Equivalence classes	286
Sample size	400

Table 1. Simulation parameters

V. Results

A first important insight from our simulations with the default setting is that 'always untruthful' does not pay off, for all equivalence classes. For all 286 equivalence classes and each utility function ($utility\_count$, $utility\_sum$, $utility\_rep$, $utility\_rank$), 'always truthful' brings at least the same or higher utility. So 'always truthful' is an equilibrium strategy (with the confidence introduced earlier and in the specific setup explicitly investigated here).

Next, we present results gained by varying settings as follows: (1) probability of misperception of the nature of comments, (2) agreement/disagreement ratio, (3) weighting function, (4) mixed strategy of synthetic participants, (4) mixed strategy of the controlled participant.

Probability of misperception of the nature of comments. Recall that raters might misperceive the nature of comments, e.g., a rater can perceive a repetition comment as an original one and vice versa. Considering the number of comments and ratings, we have come up with the following settings. We set the misperception rate to 10%, e.g., one comment out of 10 is misperceived. We set the number of ratings received per comment to 3. With these two numbers, one rating based on a misperception already is one third (33%) of the ratings received for a comment. Given the 286 classes, 'always truthful' has paid off for the following numbers of classes for each utility function ($utility\_count$, $utility\_sum$, $utility\_rep$, $utility\_rank$): 286, 269, 207, 286, respectively. Again, $utility\_count$ and $utility\_rank$ have shown to be robust towards comment misperception. From our perspective, the poor performance of the second and the third function ($utility\_sum$, $utility\_rep$) is somewhat expected. Misperception of the same comments from both synthetic participants and the controlled one have bogged down their scores in absolute values. Furthermore, resolve of repetition is affected in the case when 2 out of 3 ratings are misperceived, e.g., it is very hard to confirm the real nature of a comment which is misperceived by a majority of raters. – Note that we have set the misperception rate for ratings, not for comments. This means that all comments have the same probability to be misperceived, which is not realistic, but it is a conservative approach.

Agreement/disagreement ratio in the community. Varying the probabilities of agreement/disagreement in the community has not affected the results. We set the agreement percentage to the following values: 90, 80, 70, 50, 40, 30, 20, and 10. In all these cases, 'always truthful' maximizes the utility of the controlled participant: In all 286 equivalence classes this strategy outperforms the untruthful strategy.

Weighting function. In the setups considered so far, we calculate participant weights as the minimum of the various indicator values. In this way, we show that we deem all criteria equally important and thus, we nudge discussants to perform good regarding all of them. Nevertheless, one might think of our approach as too strict. Thus, it is interesting to check if, say, the average function (a) yields an equilibrium as well, and (b) is similarly robust as 'minimum' to other influences. Here, main insight is that all four utility functions show at least the same or higher utility for each equivalence class with 'always truthful' than with the other strategy.

Mixed strategy of synthetic participants. The rationale behind this variation has been to gain insight in the degree of trembling that is tolerated. When varying untruthfulness of the ratings of the synthetic participants, we arrive at the results in Table 2. There, the columns stand for the utility functions, and rows correspond to degrees of untruthfulness of synthetic participants. The values are the numbers of classes (out of 286) where 'always truthful' is superior.

Below:Untruthfulness(%) Right:Equivalence classes per Utility Function	utility_count	utility_sum	utility_rep	utility_rank
5	286	281	274	286
10	286	272	228	286
30	286	213	109	286

Table 2. Varying the share of untruthful ratings

We conclude that the first and the fourth utility function ($utility\_count$, $utility\_rank$) are the robust ones. Apparently, they are not affected by the untruthful behavior of other participants at all. On the other hand, the remaining utility functions ($utility\_sum$, $utility\_rep$) show deviations for equivalence classes when the number of repetitions is high, between 70% and 100%. Our explanation is that summing up absolute values of scores is not exactly helpful; the formulas are overly complex in order to really assign meaning to values. It rather is the comparison of scores that is conclusive. Regarding $utility\_rep$, with hindsight, we can say that it does not really address the primary concern of a participant who behaves strategically, which rather is pushing his opinion. Thus, while these utility functions show a weaker performance, this does not disturb us.

Mixed strategy of the controlled participant. We alter the behavior of the controlled participants by having 10% and 20% of truthfulness. While these are trembling percentages that are not negligible, the results still confirm that 'always truthful' outperforms the other strategy. For all 286 equivalence classes, 'always truthful' brings greater utility. We see this as a positive result; even in the case of mixed strategies, untruthful behavior does not pay off.

VI. Conclusions

In forum discussion, feedback of different types is crucial for a complete and comprehensive evaluation. On the other side, it gives way to participants to behave strategically in order to back up their specific interests and to push their opinion. In this article, we have studied the outcome of different rating strategies in the presence of several feedback options. Our core concern has been to investigate whether 'always truthful' as a rating strategy is an equilibrium strategy, i.e., it pays off to behave truthfully if the other participants behave truthfully as well. Next, we are interested in the question whether untruthful behavior pays off in certain cases. To address these issues, we have built a formal model that mimics the characteristics and dynamics of online discussion forums.It incorporates a sophisticated weighting and scoring scheme to assess participants and their argumentation and to test the model against different participant behavior. While we have focused on exploring this particular scheme, our approach is sufficiently modular to take other schemes into account instead. Orthogonally to this, we have proposed and evaluated different utility functions, in line with the different kinds of motivation discussants might have. On a technical level, since the number of possible states is huge, we have relied on an unbiased, representative sample of the states and have focused on the equilibrium, i.e., expected utility is maximal. As a main result, the strategy 'always truthful' is an equilibrium strategy. It also is superior to the reference strategy with some modifications of the model. An important takeaway of our work is the method itself. The setting is very complex, obvious concepts like (conventional) equilibrium are not applicable, and the number of states is huge. Nevertheless, we have proposed a respective method, which can serve as a basis to analyze settings not explicitly studied in this article.

References

[1] Arpita Ghosh and Patrick Hummel, "Implementing Optimal Outcomes in Social Computing: A Game-Theoretic Approach," vol., abs/1202.3480, 2012.

[2] Reinhard Selten, "Reexamination of the Perfectness Concept for Equilibrium Points in Extensive Games," Working Papers, 023, Aug 1974.

[3] Gnana Bharathy and Barry G. Silverman, "Validating Agent-based Social Systems Models," in *Winter Simulation Conference*,

[4] Sanja Tanasijevic and Klemens Böhm, "A New Approach to Large-Scale Deliberation," in *Cloud and Green Computing (CGC)*,

[5] Brian, Skyrms, "Dynamic Models of Deliberation and the Theory of Games," in *Proceedings of the 3rd Conference on Theoretical Aspects of Reasoning About Knowledge*, ser. TARK '90.

[6] Skyrms, Brian, "Deliberational Equilibria," vol., Volume 5, Issue 1 , pp 59-67, 1986.

[7] C. Bicchieri, "Strategic Behavior and Counterfactuals," *Synthese*, 1988.

[8] Singh, Vivek K. and Jain, Ramesh and Kankanhalli, Mohan S., "Motivating Contributors in Social Media Networks," in *Proceedings of the First SIGMM Workshop on Social Media*, ser. WSM '09.

[9] Archak, Nikolay and Sundararajan, Arun, "Optimal Design of Crowdsourcing Contests." p., 200.

[10] Yang, Jiang and Adamic, Lada A. and Ackerman, Mark S., "Crowdsourcing and Knowledge Sharing: Strategic User Behavior on Taskcn," in *Proceedings of the 9th ACM Conference on Electronic Commerce*, ser. EC '08.

[11] Dominic DiPalantino and Milan Vojnovic, "Crowdsourcing and All-pay Auctions," in *Proceedings 10th ACM Conference on Electronic Commerce (EC-2009), Stanford, California, USA, July 6–10, 2009*,