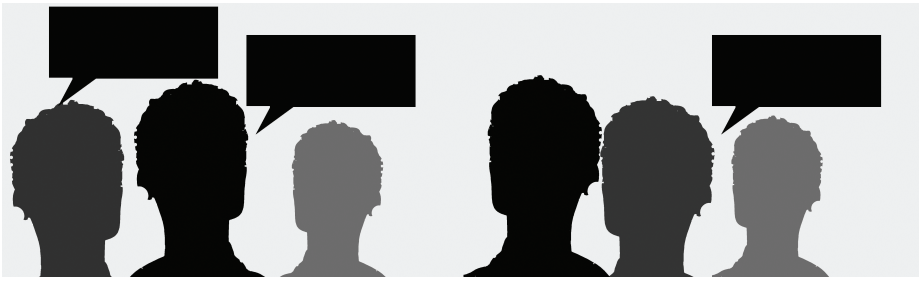
Sanja Tanasijevic

Argument-Based and Multi-faceted Rating to Support Large-Scale Deliberation



Karlsruhe Institute of Technology

Argument-Based and Multi-faceted Rating to Support Large-Scale Deliberation

zur Erlangung des akademischen Grades eines

Doktors der Ingenieurwissenschaften /
Doktors der Naturwissenschaften

der Fakultät für Informatik
des Karlsruher Instituts für Technologie (KIT)

genehmigte
Dissertation

von

Sanja Tanasijevic

aus Jagodina, Serbien

Tag der mündlichen Prüfung:	22. April 2016
Erster Gutachter:	Prof. Dr. Klemens Böhm
Zweiter Gutachter:	Prof. Dr. Karl-Martin Ehrhart

Contents

1 Deutsche Zusammenfassung

Deliberationsforen wie beispielsweise Diskussionsforen, die Argumente und Schlussfolgerungen nutzen, um zu Entscheidungen und Lösungen zu gelangen, spielen eine wichtige Rolle in der Öffentlichkeit. Methoden zur Evaluation des Inhaltes solcher Foren gewinnen immer mehr an Bedeutung. Einige bestehende Platformen für Online-Diskussionen zeigen häufige Probleme auf, wie u.a. die Wiederholung von Argumenten oder vom Thema abweichende Diskussionen. Darüber hinaus stellt der überwiegende Anteil schwer zu beurteilender Inhalte ein Problem dar. Ein vielversprechender Ansatz ist die Einführung von Feedback-Möglichkeiten unterschiedlichen Typs. Beispielsweise können Teilnehmer einen Post anhand unterschiedlicher Kriterien beurteilen, wie der Grad der Zustimmung oder Ablehnung, die Originalität oder Relevanz für das diskutierte Thema. Daher schlagen wir ein einfaches Argumentationsmodell für die Kategorisierung des Inhalts und verschiedene Bewertungstypen für die Beurteilung der Forumsqualität vor. Das Modell beinhaltet das folgende Evaluierungsschema:

(1) Benutzer erhalten Wertungen basierend auf diversen Kriterien wie bspw. Originalität ihrer Posts anhand des Feedbacks anderer Benutzer;

(2) Posts werden anhand der Rate von Zustimmungen/Ablehnungen die sie erhalten haben, sowie anhand der Wertungen von Bewertern und Autoren beurteilt.

Zur Evaluation unseres Argumentationsmodells haben wir ein Experiment durchgeführt, das nah an reale Bedingungen mit mehr als 100 Teilnehmern ist. Studenten des Datenbank-Kurses unserer Universität diskutierten verschiedene für sie relevante Themen. Die anschließende Studie und umfangreiche Auswertung der Versuchsdaten erbrachten den Beweis, dass die Teilnehmer unser Deliberationsmodell gut akzeptierten.

Die nächste Phase untersuchte die verschiedenen Bewertungsstrategien von Teilnehmern, da das Feedback der Teilnehmer der wesentliche Punkt unseres Evaluierungsansatzes darstellt. Wie in vielen anderen Projekten die sich mit unterschiedlichen Verhaltensstrategien und den zugrundeliegenden Nutzenfunktionen beschäftigen, haben wir ein formelles Modell basierend auf Konzepten der Spieltheorie und, im Besonderen, nicht-kooperativer Spiele entwickelt. Mittels formeller Analyse untersuchen wir die Effekte auf verschiedene Verhaltensstrategien und deren Einfluss auf den Evaluationsprozess. Das formelle Modell lässt uns folgende Fragen betrachten: Wann lohnt sich unehrliches Bewertungsverhalten und ist ehrliches Verhalten eine Gleichgewichtsstrategie? Wir schlussfolgern, dass die Strategie, Posts immer ehrlich zu bewerten, eine Gleichgewichtsstrategie ist. Darüber hinaus hat unser Evaluierungsschema bewiesen, dass es in vielen Fällen robust gegen unehrliches Verhalten von Teilnehmern ist.

Unsere Arbeit adressiert generelle Schwierigkeiten, die Online-Diskussionen in vielen Applikationen haben. Die Arbeit selbst bietet eine umfangreiche Betrachtung eines komplexen Problems. Die Evaluation unseres Ansatzes beinhaltete experimentelle Methoden unterstützt durch Studienergebnisse und experimentelle Datenanalyse. Zusätzlich verwendeten wir eine formelle Analyse, um weitere Erkenntnisse und Bestätigungen für unser Forum-Modell zu sammeln.

Der vorgestellte Ansatz zeigt sich sehr vielversprechend in Bezug auf die Organisation konstruktiver Online-Diskussionen basierend auf der Argumentation und Schlussfolgerung

sowie der Evaluation der Beiträge basierend auf der Diskussion selbst und ihrer Struktur. Das vorgestellte Modell wurde von den Teilnehmern sehr gut akzeptiert und zeigt sich in den meisten Fällen robust gegenüber unehrlichem oder strategischem Verhalten.

2 Acknowledgments

This thesis has been completed as a joint work of me and my two supervisors, professor Klemens Böhm and professor Karl-Martin Ehrhart. I would like to thank professor Klemens Böhm for his tremendous support, not only technical but also as supervisor. He had always encouraged me and it has really helped me to overcome some stressful moments during my PhD studies. Professor Karl-Martin Ehrhart has contributed significantly with his suggestions and comments especially for our third paper and game theory related issues.

Next, I would like to thank all my institute colleagues for all valuable inputs and discussions. Pavel Efros and Hoang Vu have given me useful insights while reviewing my papers. Muhannad Ali was shortly my office mate and he has been not only a great friend but he has also contributed on the presentation level during my thesis write up. Finally, Fabian Keller has shared some important information regarding the thesis defense.

My personal thanks go to Fabrizio Massaro for all the happy moments that gave me energy and inspiration to wrap up my work. Thank to all my friends (Aleksandar Papic, Koviljka Zecevic, Miodrag Tezic, Chunyan Li, Caslav Bozic) for being great friends and honest supporters. Special thanks to my friend Bane Timotijevic, who has not only been my greatest inspiration for starting my PhD studies, but he has also been the source of immense motivation, help and support along the way.

Last but not the least, I would like to save my biggest thanks to my twin sister, Tanja, who has been a great encouragement and motivation along these long three years of my PhD studies. She has believed in me and my decisions even more than I ever did and I will be forever thankful for that and moreover for having her by my side.

3 Abstract

Deliberation forums, e.g. discussion forums that use arguments and reasoning to come 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. Some existing platforms for online discussions show common problems such as repetition of arguments or off-topic discussions etc. Furthermore, there is a problem of overwhelming amount of content that is hard to assess. 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 scales of agreement/disagreement, or the originality or relevance to the topic currently discussed. Thus, we propose a simple argumentation model to categorize content and differentrating types to assess the quality. The model incorporates the following evaluation scheme:

(1) Users earn weights based on several criteria, such as originality of their posts according to feedback by others;

(2) Posts are appraised based on the rate of agreement/disagreement feedback they have obtained and the weights of raters and authors.

To evaluate our argumentation model, we conducted an experiment that is close to a real-world conditions with more than 100 participants. Students of the database course at our university discussed various topics relevant to them. The subsequent survey and extensive analysis of experimental data offer proof that participants accepted our deliberation model well.

The next phase explores different rating strategies of participants since feedback of participants represents the crucial point of our evaluation approach. As in many other projects dealing with various behavioral strategies and underlying utility functions of the participants, we have developed a formal model using the concepts of game theory and, more specifically, non-cooperative games. By means of formal analysis we examine effects of different behavioral strategies and their influence on the evaluation process. The formal model lets us study the following questions: When exactly does untruthful rating behavior pay off, and is truthful behavior an equilibrium strategy? We conclude that the strategy 'rate posts always truthfully' is an equilibrium strategy. Furthermore, our evaluation scheme proved to be robust towards untruthful behavior of participants in many cases.

Our work aims at addressing common difficulties online discussions face in many applications. The work itself presents comprehensive consideration of a complex problem. Evaluation of our approach has included experimental method, supported up with survey results and experimental data analysis. Additionally, we have employed formal analysis to gain further insights and verifications of our forum model.

The proposed approach has shown to be very promising in terms of organizing constructive online discussion based on argumentation and reasoning and evaluating contributions based on the discussion itself and its structure. The proposed model has been very well accepted by participants and has showed to be robust towards untruthful or strategic behavior in the most cases.

1 Introduction

Deliberation is the act where communities identify possible solutions for a problem and the one(s) from this space that best meet their needs [1, 2]. The spectrum of communities whose discussions rely on reasons and arguments is broad: It not only includes groups of citizens, from (small) municipalities to much larger administrative units. It also ranges from communities in science and technology, including the teams developing software, and communities of online gamers to groups of experts within large companies or organizations. Many communities are small, consisting of about, say, 100 or 200 individuals.

The question how communities can come to decisions and solutions that are satisfying for most of their members continues to be fundamentally important. There currently are various experiments and online projects trying to foster deliberation. To give some examples we list following projects: discussions and voting on budget planning for the German cities of Essen or Stuttgart (essen-kriegt-die-kurve.de, buergerhaushalt-stuttgart.de). The limitations of these two projects, however, are exemplary of the ones of many other initiatives. For instance, essen-kriegt-die-kurve lets individuals propose concrete budget cuts and discuss these proposals. This project 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. With Bürgerhaushalt Stuttgart in turn, individuals can come up with proposals others can then vote on without bringing any reasoning or arguments.

In practice, deliberation faces problems: Online discussions sometimes generate poorly organized, unsystematic and redundant contributions of varying quality [3]. Significant effort is required to extract important issues, ideas and arguments. Thus, to address these various challenges, it is mandatory to radically reduce redundancy and encourage clarity. Some recent projects, e.g., Deliberatorium [4], have tried to apply a very formal argumentation model to bring structure to online discussions and to facilitate content evaluation. However, such rigid formalisms often undermine the natural discussion flow and require a lot of effort from participants.

Thus, we have proposed a relatively simple argumentation model to categorize content and different rating types to assess its quality. The rationale has been to give a clear structure to the discussion and to nudge discussants towards deliberation. In more detail, participants discuss different proposals by posting arguments in favor or against these proposals). In that way, participants also categorize their comments based on its content; examples of respective comment types are pro argument or contra argument. On the other hand, a participant can post a feedback of different types for comments of other participants. A feedback can be related to the argumentation presented, but it might also refer to the clarity

of writing, to the tonality of comments, or to the types of the comments. Basically, the evaluation should single out comments which are not related to the topic of the discussion or repetitions of previous comments or that are offensive or provocative. At the same time, and mainly orthogonally to these dimensions, it is of course interesting to know whether a community agrees or disagrees with a certain comment. One-dimensional feedback options such as 'thumbs up', 'like' buttons etc. are too undifferentiated to this end.

In the next step, we 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 in the discussion. Next, we propose a scoring scheme for evaluating comments as arguments. 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. In our setup, proposals are alternatives to each other, and they are also evaluated using a score scheme that is argument-based. Based on all this information, our approach assesses potential solutions to discussion subjects which participants have proposed in the course of the discussion. With our approach, collecting ideas for solutions is as important as their evaluation. This is in slight contrast to other recent deliberation projects such as ConsiderIt [5], which focuses on the collection of pro and contra arguments. In real life and in other studies, e.g., ConsiderIt [5], Deliberatorium [4], one usually takes pro and contra arguments into account when making a decision. Similarly, for each proposal we collect pro and contra arguments and evaluate them using scoring scheme which is argument-based. In other words, decision-making is mainly based on the structure and assessment of the arguments, as opposed to voting.

As presented, participants are seamless part of the evaluation process. Nevertheless, participants in discussions have different interests, which motivate them to issue feedback on posts. In the presence of feedback of different types, they can behave strategically in order to back up their specific interests and to push their opinion, cf. [6]. Thus, we aimed at studying this specific setting, namely *various rating strategies* in the presence of *feedback of different types*. The core questions are when exactly untruthful rating behavior may pay off, and how this can be avoided. Additionally, it might seem worthwhile to examine whether truthful behavior is an equilibrium strategy. In other words, truthful behavior represents a steady point of the game and no participant has interest to deviate from this behavior.

To evaluate our approach we have conducted an experiment that is very close to the real setting, with more than 100 participants. Students of a database course at our university have deliberated on various topics relevant to them. In a subsequent survey and by means of extensive analysis of experimental data we gained proofs that participants accepted our deliberation model very well. The participation was high. The discussants expressed their satisfaction with the evaluation approach that we have used and as well as with the outcome decisions.

Step further was to look at different rating strategies of participants since feedback of participant represents a crucial point of our evaluation approach. As in many other project dealing with various strategies and underlying utility functions of the participants we have developed a formal model using the concepts of non-cooperative games. By means of formal

analysis we have examined effects of different behavioral strategies and their influence on the evaluation.

Outline of the thesis: The thesis represents a work done in three years of extensive development and design of a concept for the forum model that will successfully cope with the common issues of online discussions. We have started with the introduction of our research problem and motivation for our work. We continue with the related work presented in the Chapter 2. Considering interdisciplinary approach of our work, the chapter gives a broad overview of related research, projects and neighboring fields. The chapter itself has several sections focused on various topics related to our work. Chapters 3 and 4 introduce in more details the fields tightly connected to our research. Formal and informal logic, argumentation, argumentation theory are subjects of the chapter 3. Game theory basic concepts and formalization are described in the chapter 4. The outcomes of our work are presented in chapters 5 and 6 as follows. First, we introduce our deliberation forum model starting from requirements, design decisions, and formalization details. We describe the experiment we have conducted and a subsequent survey and wrap up with the results gained from the participants' reports and the comprehensive data analysis. Next, in the chapter 6 we introduce the second part of our work, namely, formal analysis of our forum. We present the design and implementation details of the formal model. We show the results of simulating different behavioral strategies on this formal model incorporating the proposed evaluation approach and different types of feedback. In the last chapter we conclude with a summary and possible research directions for a future work.

2 Related Work

In the following chapter, we are going to give an overview of related research and projects. Considering the multidisciplinary approach of our work the range of different projects, we will present, is quite broad. Thus, we have divided the related work in four separate sections.

We start with presenting the research in the field of e-democracy. We will introduce some important related projects and experiments. Usage of information technologies has opened new prospects in developing democracy by increasing public informedness and engagement in discussions and decision-making. At the same time various issues and challenges have arisen. The underlying idea of our work was to support communities to conduct effective discussions, be they students groups, university boards, municipalities etc.

In the next section, we present a research field dealing with arguments, argumentation theory, systems for argumentation assistance and analysis. We outline some important works related to argumentation theory and review different tools based on these theories. Various approaches propose different ways of handling arguments in everyday discourse which are neither deductive nor inductive. Nevertheless, the majority of these approaches appear to be overly complex and still incomplete. Hence, we have decided for rather simple than exhaustive argumentation model that will facilitate deliberation and bring some structure in the discussion without necessarily disturbing the discussion flow with strict formalism.

Subsequently, we introduce various related projects and experiments that study problem of motivation of participants to contribute and actively participate in online platforms. This issue has appeared to be very important for developing online communities, thus, significant efforts are invested to motivate and nudge discussants to active participation by means of various rewarding and reputation system. In our settings, active participation is crucial not only for collecting arguments but also for receiving users' ratings which are the base of our evaluation approach.

Following is the section about decision-making. Process of decision-making especially in the context of online platforms continues to be very interesting for researchers. Information technologies have certainly had a strong impact on group decision-making bringing people virtually together with the possibility to exchange information and share ideas more easily, and consequently, come to decisions of higher quality. The underlying premise of our work was to design a forum model that will facilitate group decision-making and yield decision with higher acceptance by a community.

In the end, we present research in the field of formal modeling and game-theory. There are some interesting related projects employing formal modeling and game-theoretic approach in the analysis of different social systems such as deliberation or crowdsourcing projects.

A common ground of our work and these projects is usage of formal modeling and game-theoretic analysis to study the effect of different behavioral strategies in the deliberation forum with different types of feedback.

2.1 E-democracy

Literature on e-democracy has emerged in the early 1990s. "Online forums are a development of historic significance, for there has been practically no innovation in many-to-many communication in over two thousand years" [7]. Internet has a novel possibility of virtual gathering of large number of people geographically scattered quickly and cheaply. The true premise of the Internet lies not merely in its ability to bring large number of people into "one room" but in its ability to structure the room in ways that no room can be structured [8]. Nevertheless, not all online forum discussions understand deliberation. The strengths of deliberation are idea synergy and diversity, results checked by many and collective wisdom. Deliberation is an ideal form of discussion in which participants share their considerations in order to make decisions of higher quality and democratic legitimacy [9, 10, 11, 12, 13, 14, 15, 16]. An important impetus behind the growing interest in citizen deliberation has been the claim that immersing citizens in deliberative contexts can help educate them about political issues, countering the low levels of political knowledge and sophistication in the mass public [14].

Research has shown that the public appears to be uniformed, misinformed, and neither very knowledgeable nor sophisticated about politics [17, 18, 19]. Low levels of knowledge need to be addressed because knowledge matters. Authors[18] found that general political knowledge, measured as a series of factual items, has positive and significant effects on political tolerance, electoral participation, and the fact whether citizens hold any opinion in the first place. Similarly, [20] has found that in panel studies, political knowledge significantly influences response variability, perhaps because those with greater knowledge are less likely to hold "non-attitudes" [17]. Even if deliberation can enhance citizens' political knowledge and sophistication, it also faces important equality concerns. Deliberative knowledge benefits might support inequitably, much as in traditional political participation [21, 22]. Of course, deliberative democratic theorists recognize this problem and consequently some advocate for economic and institutional reforms to accompany the expansion of deliberative practices (e.g., [23, 24]). In addition, even if deliberation does not worsen inequality, inequality, nevertheless, presents a challenge because deliberation would ideally proceed with all citizens equally able to represent their viewpoints [25].

Jayasena & Karunaratna defined e-democracy as the utilization of information and communication technologies (ICTs) for enhancing a countrys democratic processes and empowering its citizens [26]. E-democracy consists of all electronic means of communication which enable or empower citizens in their efforts to hold rulers/politicians accountable for their actions in the public realm [27]. This enables citizens to effectively channel their complains and opinions to governmental bodies about issues concerning the whole society [28]. Some authors refer to these electronic channels of communication as the cyberspace, seen as a vital channel in civil society in which individuals and groups can become informed about

issues, discuss and debate these issues autonomously, and ultimately have an impact on policy agendas [29].

E-participation grew out of the field or is a subset of e-democracy which seeks to empower people with the help of information and communication technologies to enable them to integrate in bottom-up decision-making processes, and to develop social and political responsibilities [30]. In the article [31] it has been presented an overview of the state-of-the-art research in this field. The figure below depicts different interested parties and different forms of e-projects:

- E-voting
- E-negotiation
- E-deliberation
- E-petitions
- E-campaigning

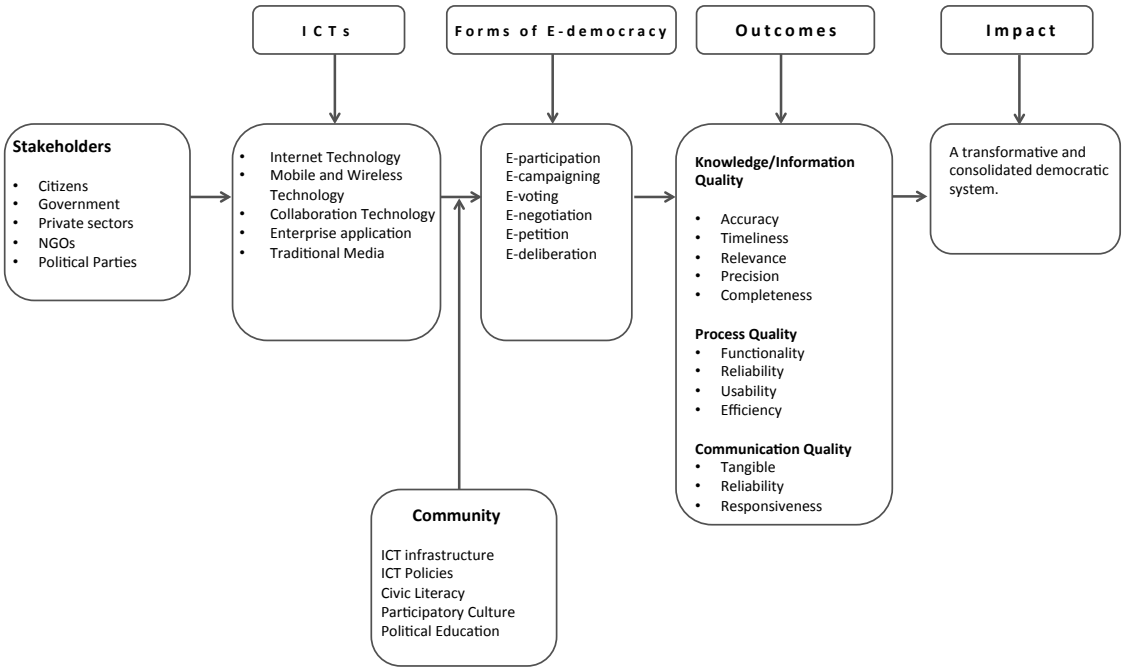


Figure 2.1: E-democracy chart

These different types of projects have different focus i.e. collecting votes from participant as a common democratic practice (e-voting); mobilizing and gaining support for different projects (e-campaign); organizing public initiatives in the form of petitions (e-petitions). Finally,

e-deliberation projects represent another group of projects aimed at fostering interactive communication on political issues in order to strengthen the legitimacy of decision [32] and increase the awareness and informedness of the public. See Figure 2.1.

There are several problems with deliberation caused by nature of the participants such as the diversity of their views [33], willingness and tendency of respecting deliberative rules [34] or the nature of content allowed in deliberation etc. Still the importance of deliberation and reflective consideration is very crucial in the community discussions. Gastil defines deliberation in terms of five practices: (1) acquisition of information; (2) identification and prioritization of values; (3) consideration of broad range of solutions that might address a problem; (4) reflection of potential consequences and trade-offs entailed by solutions; and choosing among proposed solutions [35]. Various projects have examined the possibilities of nudging discussants in the direction of the more balanced considerations that include reflexive examination of own reasoning as well as reasoning of other viewpoints. Furthermore, effective moderation is considered to be very important since the perceived anonymity in online forums weakens the norms of constitutive/self-censorship that regulate face-to-face behavior. It is thought that this can lead to "flame wars", polarized debates and dominant minorities. Thus, while the anonymity of online environments may diminish the psychological thresholds that can limit participation, it may also exacerbate them by inhibiting the social cooperation needed to accomplish complex communicative tasks [36].

Deliberation has proved to significantly increase participation and engagement in an experiment to match face-to-face deliberation of random sample of citizens discussing policy issues and electoral choices [37]. An interesting project aimed at online deliberation was carried out in the community of the Dutsch city of Hoogeveen. The experiment was conducted before the municipal elections and the central issue was how the political discussions contribute to the creation of the public sphere [38]. However, the project focus was on the three dimensions of behavior in online discussions: indications of special treatment based on the status of participants, indications of intimidating or denigrating language and indications of domination within a discussion. Agora project [39] has undertaken the research of open-source freeware and prospect of online deliberation as a conflict between intensified grouping, opinion polarization, and normative conformity of the group as opposed to deliberation of citizens that should avoid poor transmission of social cues and disinhibition [40]. Agora project was a part of the broader initiative so-called VAProject financed by the Government with aim to contribute to deliberative democracy research. The main two goals of the project were: (1) to develop of an open-source freeware for democratic research and (2) to conduct research on democratic deliberation. Additionally, project aimed at isolating the factors that encourage political deliberation and applying IT concepts. Positive outcomes can be categorized in two groups: decision quality effects and community effects. Application of IT in political decision-making may have positive or negative effects on the quality of decisions. To assess the quality of decisions following measures are proposed: conformity of decision-making, rationality of choice, correspondence of decisions with respondent values, identification of common good etc. On the other side, community effects correspond to feeling of citizenship and the well being of the community. Community effects examined include: the effects of deliberation on the community-mindedness of participants, degree of conflict, capacity of deliberation to build social networks and social trust etc. [41, 42]. The project itself was very important because it was one of the first project to use information technology for high-quality political discussion and to identify factors that encourage it.

E-Liberate project pulls out further difficulties in organizing online deliberation to provide and support it optimally [43]. So far there are limited number of applications available. It is difficult to design and implement them in the situation when economic interest is not that obvious. Next, the process of deliberation is difficult due to the complexity of content and potential contributors perceive the payoff far below the efforts invested. Finally, no matter the increasing role of online deliberation in public and politic life, in most societies it has limited influence. Nevertheless the project aimed to facilitate decision-making in communities in equitable and collective manner by employing Roberts rules of Order. Beginning in the late 1800s, Henry Robert devoted over forty years to the development of Roberts Rules of Order, a set of directives that designated an orderly process for equitable decision making in face-to-face meetings. One of the most important design objectives was to guarantee every attendees opportunity to make his or her ideas heard while ensuring that the minority could not prevent the majority from making decisions. The system uses simple and straightforward interface for real-time meetings as well as offline meetings. The contributions of online deliberations are typed to allow automated management of the interactions. Nevertheless, strict rules on participation and presence in the discussion appear to be hardly obtainable in online environment.

Previously, we have given an overview of research and experiments in the field of e-democracy. These various works focus on democratic potential of online deliberation. Some of them have conducted experiments under controlled conditions (online and offline users, trained groups, control and experimental groups, sampling of participants) etc. The variety and number of the similar project is constantly growing. Still, not that many freewares are available allowing organization of a wider e-democracy initiative. This is due to the complexity of the discussed issues and also specifics of settings intended to involve larger community participation with different level of education, communication skills and information.

2.2 Argumentation schemes

Deliberation has shown to have positive effects in many different contexts. In practice however, deliberation faces serious challenges, including disorganized, redundant content, quantity over depth, strong polarization, and dysfunctional arguments [4]. Large-scale argumentation systems claim to address these shortcomings, by providing a systematic structure that radically reduces redundancy and encourages clarity [4]. Respective projects structure content by means of different argumentation models.

Argumentation schemes capture stereotypical patterns of human reasoning, especially defeasible ones like argument from expert opinion, that are proven as troublesome to be viewed deductively or inductively [44]. Argumentation schemes have more recently been identified in computation domains to potentially bring significant improvements in reasoning and communication. Building formal system based on argumentation schemes and expressed in a natural language has widen the scope of their usage and computation efficiency. The aims of formalization should: (a) remain sufficiently close to linguistic practice in terms of richness and flexibility of natural argumentation; (b) render a model which is understandable [45].

Literature has proposed various argumentation schemes that represent inferential structures of arguments used in everyday discourse. There are many theoretical and practical works starting back in ancient Greece and deductivism of Aristotel. Perelman and Olbrechts-Tyteca identified and defined many distinctive kinds of arguments [46]. Arthur Hastings Ph.D. thesis made even more systematic classification by listing many of these schemes and sets of critical questions matching each one of them [47]. Toulmin extended the traditional *premise-claim* model with additional elements: *warrant*, *backing*, *qualifier* and *rebuttal* [48]. In his work, Toulmin focused on so-called *practical arguments* or *substantial arguments* having the justificatory function of argumentation, as opposed to the inferential function of theoretical arguments. While theoretical arguments make inferences based on a set of principles to arrive at a claim, practical arguments first find a claim of interest, and then provide justification for it. Verheij gives a formal elaboration of Toulmin's scheme and extends it with evaluation of Toulmin's arguments [49]. Although it is impossible to retain all of Toulmin's ideas and implement them in full form, the starting point of the reconstruction of Toulmin's scheme is a theory of dialectical argumentation, so called *DefLog* [50]. When an assumption is *prima facie* (justified at first sight), there can for instance be a reason against it. Then the assumption is not actually justified but rather said to be defeated [51].

Verheij has extended his interests to argumentation assistance systems. Note that *argument assistance systems* distinguish from the more common *automated reasoning systems*. The latter automatically perform reasoning on the basis of the information in their 'knowledge base' and thus can do (often complex) reasoning tasks for the user. On the other side, argument assistance systems do not (or not primarily) reason themselves; the goal of assistance systems is not to replace the user's reasoning, but to assist the user in his reasoning process.

The first argument assistant system proposed by Verheij was *ARGUE!* and it was based on the logical system *CUMULA* that abstractly modeled defeasible argumentation [52]. In *CUMULA*, a procedural model of argumentation with arguments and counterarguments (in the sense of trees of reasons and conclusions) can be defeated. The defeat of arguments results from attack by other arguments, as expressed by defeaters. A defeater indicates which set of arguments attacks which other set of arguments. *CUMULA*'s defeaters allow the representation of several types of defeat, including defeat by parallel strengthening and by sequential weakening [52]. Nevertheless, the system was not sufficiently natural for the representation of real-life argumentation and it appeared to be too complex for the intended users. The project was mainly interesting from a research perspective, as a realization and verification of a particular theory of defeasible argumentation [51]. *ARGUMED* was a second argument assistant system that featured graphical representation of arguments and evaluation of the statements involved. The development of argument assistance systems is still in experimental phase due to many difficulties such as the lack of a canonical theory of defeasible argumentation, special requirements for the design of user interfaces so that arguments can be sensibly and clearly presented to the users [51].

In recent years, numerous tools were designed for argument assistance and argumentation analysis. However, all these argumentation models proposed by various authors have the problem of completeness. A completely systematic justification of defeasible argument schemes is ruled out by their non-monotonicity and the contextual determination of their

acceptability [53]. Defeasible argumentation schemes need practical justification that solves three problems: (1) how to classify arguments into different types, (2) how to identify arguments and conclusions, (3) how to formulate critical questions used to evaluate arguments [44]. One of the good examples is *Araucaria* developed by Reed and Rowe [54]. The assumptions behind it is that a single text might be analysed in several different ways, depending upon a variety of analytical choices. The text is analyzed to produce a diagram with the links between original text and the corresponding components. *Araucaria* identifies propositions as vertices in a diagram. The system provides an interface which facilitate diagramming of arguments and saving it using *AML* (the argument markup language) designed in XML. *AML* defines a set of tags that indicate delimitation of argument components (loosely, propositions), tags that indicate support relationships between those components, and tags that indicate the extent of instances of argumentation schemes. Both *Araucaria*, and the markup language in which analyses are saved, exploit the typical tree structure of argumentation. *Araucaria* allows a conclusion to be expanded to show its supporting premises, and for each of those premises supporting arguments can be seen, and so on.

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 [4]. *Deliberatorium* is a large-scale argumentation system. It introduces the concept to overcome the limitations and emphasize the potential of social media for better decision-making. 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) [4]. Deliberation map improves the logical order of argument relying on visual effect of a tree structure of arguments. Users are encouraged to disagree by posting new posts or collaboratively refine proposed solution ideas. Thus, the map represents the whole design space of possible solutions or ideas that can be refined, combined or compared on equal basis. If one disagrees with an idea or argument, the user should not change that post to undermine it, but should rather create new posts that present the strongest ideas and arguments they can muster, so all contributions can compete on an even basis. Moderators help the instructions are followed and respected. A screenshot from *Deliberatorium* can be seen in Figure 2.2. Deliberation map was used in several communities such as master students at University of Naples, in Intel Corporation and it showed promising results especially in large communities. Considering the costs of building a map, it is expected that at some point benefit exceed costs of individuals. Nevertheless, the strong formalism undermine the natural flow of discussion which is seen as a main disadvantage. Thus, it cannot be guaranteed that it will be accepted in a wider community (public usage).

IBIS argumentation formalism has been applied successfully in hundreds of collective decision-making contexts. Rittel and Kunz describe IBIS as an concept for information systems "to support coordination and planning of political decision processes" [55]. The approach has been applied for design and planning and thus it is considered to be "argumentative" and "discursive". The formalism brings structure by introducing following elements: (1) *issues*; (2) *topics*; (3) *positions* and (4) *arguments*. Issues are main elements around which the discussion is developed. Different issues can be classified under one topic. Participants take different positions, defend their position, oppose to other in the form of arguments. Further important points are relations that build the actual structure by connecting arguments to

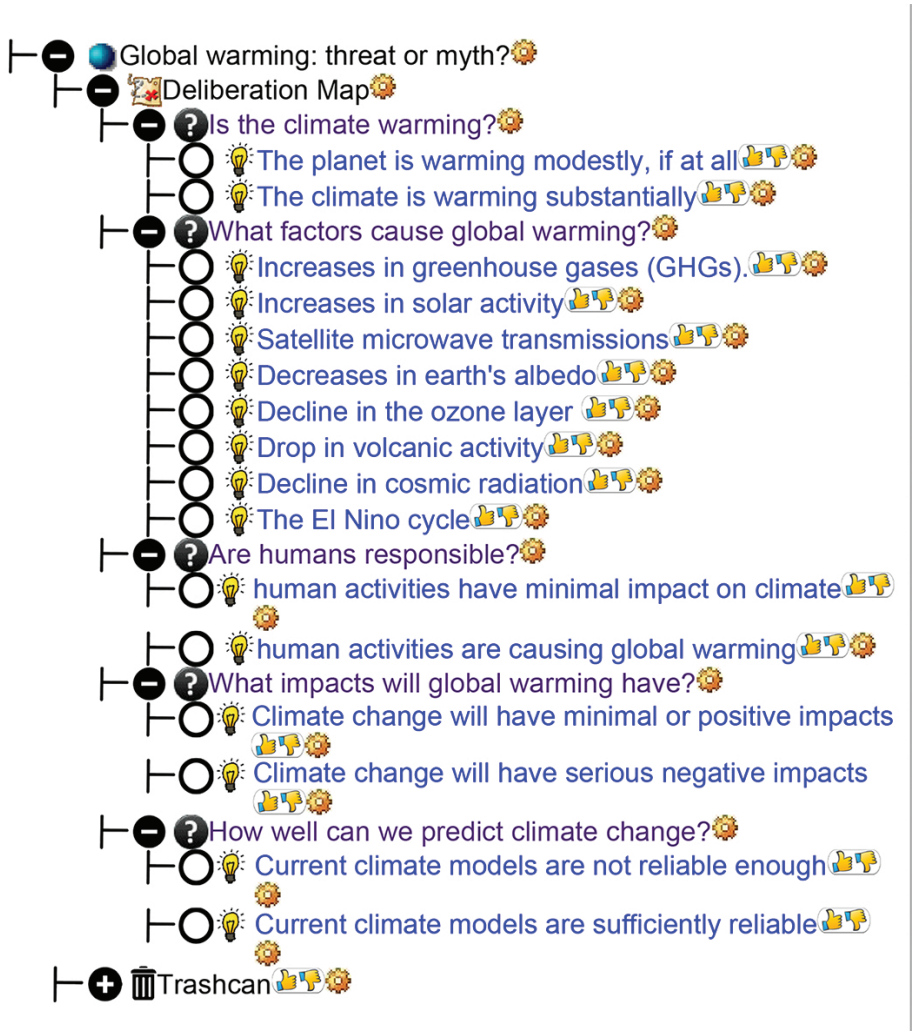


Figure 2.2: Screenshot from Deliberatorium

positions, positions to issues, issues to topic. In the end, operations relates to handling with issues such as: raising an issue, activating external knowledge sources, creating relations etc. IBIS systems have been applied in engineering, architectural design and numerous other project. In our context interesting project to mention is *Cohere* [56]. *Cohere* is a web tool for social bookmarking, idea-linking and argument visualization that employs IBIS formalism. It incorporates the Web 2.0 principles and at the same time support argument analysis and visualization. The project introduces following constituents derived from IBIS: *Ideas*, *Questions*, *Answers*, *Pros* and *Cons*. *Cohere* provides a platform for collaborative deliberation and debate mapping over the internet, with primarily asynchronous use. Real time mapping requires a separate tool for user interface optimization for rapid mapping. Furthermore, Waltons argumentation schemes and associated Critical Questions, rendered as XML files can be modeled in Cohere.

A fundamental prerequisite for the application of IBIS-like methods is a general acceptance of argumentation and discourse and belonging formalism. 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 [57].

Rarely, argumentation systems have also been used to enable distributed deliberations over the Internet [58, 59, 60, 61, 62, 63]. These maps tend to be poorly structured, however, because many users are not skilled argument mappers, and the scale of participation has been small. Thus, the implementation in the wider public usage can face difficulties.

2.3 Related projects

Deliberation does not necessarily occur only in discussion forums, it can also take place on websites and portals that support posting comments or feedback. Some recent projects give a broader views on issues be they discussion issues, technical issues, political or technology news by including users as active participants. *NewsCube* is a novel internet news service aimed to tackle the news bias [64]. Offering plurality in viewpoints and possibility to create users' own new or re-stated viewpoints the work aims at higher understanding from reader by giving multiple classified viewpoints. News suffer from subjectivity of authors, owners or advertisers. The core is aspect-level browsing, collecting different news articles from various sources or recommending article with contrasting or similar aspects. In order to mitigate news bias, presentations should actually lead users to read diverse aspects of a news event. Classification of aspects is central to support aspect-level browsing via clustering since it is difficult to impossible to predict and construct predefined categories.

Though more and more of our communication is taking place on the web, our web interfaces have typically not supported back-channels for others to demonstrate evidence of understanding. On the other hand, we are often invited to hit a "like" or "thumbs up" button to signal some affinity without claiming a ground. The project named *Reflect* supports active listening and grounding on the Web through restatement [65]. Reflect proposes an interface for online public discussions by engaging and motivating discussants to help facilitate large-scale deliberation with restating, finding and sharing common grounds. The authors design interfaces that help nudge people toward more reflective interactions by emphasizing the common experience of listening. The project outlines the three important points:

- improving communication satisfaction and willingness to participate by proving discussants that they are listened
- emphasizing listening skill to guide the discussion and demonstrate their viewpoints
- help other discussants to improve understanding

Reflect strongly suggests restatement as a primary mode of participation. Every comment is accompanied by a list of concise summary bullet points added by other readers/discussants. The original commenter can respond to each bullet point to make sure that the summarized points accurately portrays the point they were trying to make. Future readers are able to observe this process (and eventually participate), learning more about what was being said and how other people interpreted what was said.

Opinion Space introduces an online interface incorporating ideas of deliberative polling and collaborative filtering for visualization and navigation through diversity of comments [66]. "Participatory culture" has increased significantly the number of comments for news, blogs, videos, commerce etc. Along with this trend new issues have arisen such as: "flame wars", overwhelming amount of data and polarization of communities on websites. To cope with this, *Opinion Space* sorts out opinions to a set of controversial statements as scalar values on a continuous scale (from strongly disagree to strongly agree) placing participants by their opinion. Viewpoints are distributed taking into account the distribution of opinion and participants are encouraged to take balanced position. *Opinion Space* appears as a self-organized system that rewards participants who consider opinion of their opponents building participants' connections based on their opinion.

The importance of reflective and deliberative discussion has been demonstrated in the *ConsiderIt* project which encourages the discussants to express pros/cons points which can be created, shared and adopted by various participants [67]. The focus is on personal deliberation rather than on direct discussion with others. Thus, the software system grounds on two facts: (1) it supports personal deliberation by crafting a position; and (2) it explores aggregated positions to understand the considerations accepted by the most users. *ConsiderIt* facilitates personal deliberation supported by the considerations others are making. In that way, the system helps users explore and accept public thought derived from these personal deliberations. Participants are invited to vote for a solution problem choosing pro or contra points either from others or create/restate their own. A slider that visualize a participant's standing point can be also, manually adjusted. Evaluation of the pro and contra points is based on the frequency of their usage by participants and their opinion is then calculated as a ratio between pro and contra points.

The focal point of presented projects is to implement deliberative principle in different usage scenario such as: to mitigate effects of news bias, increase the understanding of reader, support self-deliberation etc. The aim of our project is to propose the discussion forum model built around deliberating principle employing at the same time comprehensive evaluation approach aimed at extracting valuable arguments, comparing different solutions to proposing a final solution.

2.4 Online participation

Encouraging participation is one of the greatest challenges for any online community provider. There is a large amount of literature demonstrating ways in which online communities can be effectively built [68, 69, 70, 71]. However, an online community can have right tools, right chat platform and right ethos, but if community members are not active the community will not thrive. Only a small percentage of users contribute content, only a small proportion of them contribute significant amounts of content [72, 73]. Even successful communities like Wikipedia face problems such as a large backlog of pending tasks, low retention of new editors [74]. Encouraging members to participate requires to understand mechanisms to improve and motivate participation.

Bishop has proposed a theoretical framework for understanding which factors encourage visitors to participate in online discussion forums [75]. The framework has three levels: (1)

a desire to carry out an action, such as solving a problem; (2) interpret whether taking this action is consistent with their goals, plans, values, beliefs and interests; (3) use their abilities to carry out the action and perceive themselves as a part of the group. On the other hand, authors in their work [76] have completed an experimental study with 5733 online participant. They distinguished a few categories of online communities:

- Person-oriented communities. Communities where social interactions between individuals are in focus. Examples are MySpace, Facebook, Friendster, Bebo, Orkut, Windows LiveSpace, and Hi5.
- Professional communities. Communities that focus on business networking such as: LinkedIn and itLinkz.
- Media-oriented communities. Communities that focus on the distribution and consumption of user-generated multi-media content, such as video, music or photos. Examples are YouTube and Flickr.
- Virtual-world communities. Communities that are essentially a 3-D virtual world, built and owned by their residents (the users). A typical example is SecondLife and World of Warcraft.
- Mobile communities. Communities that make it possible to have direct and indirect contact with community friends and allows users to make updates on the move i.e. Twitter.

The study based on a questionnaire has given some useful insights into age, gender and education structure for each of different platforms as well as the usage model of the belonging participants. Large communities such as YouTube and Wikipedia, are a new mean of creation and sharing of media, but the basic consumption paradigm is still to broadcast - from few to many. This is contrary to what we see in the smaller online communities where social networking is a prime focus, and content is produced by few - and consumed by few. Earlier research confirms this hypothesis showing that lurking levels are higher in some communities than others, and that lurking may vary in relation to other community variables such as size of the community, frequency of posting, and number of single messages (e.g. [77]).

Authors have listed 14 different usage patterns based on the reports and three most common once were: (1) checking if somebody has tried to contact them; (2) writing email or messages; and (3) contacting others. Based on these usage patters users are categorized in five different groups: (1) sporadics; (2) time-killers; (3)socializers; (4) debaters; and (5) actives. Ludford et al. carried out a field experiment with different methods to spark the community participation [78]. They used an existing movie recommendation website, MovieLens, and manipulated two factors: (1) similarity in ratings of the group members' movie ratings and (2) uniqueness - the raters were told that their ratings were unique in the group. Both factors have positively influenced participation. Thus, forming groups with diverse perspectives and showing people their unique qualities relative to a topic increase contributions. The same website is used in another work [79] which investigated the motivation of online users, types of the roles they take in a study that lasted over a year and had included 4000 new users. They discovered that "general volunteer motivations, pro-social behavioral history, and community-specific motivations predicted both the amount of use and specific types of activities of users engaged in after joining the community".

Consistent with functional theorizing, they found that different motivations, and different histories of pro-social behavior, led to different patterns of behavior.

Another group of projects have explored reputation mechanisms, their connection with participation and user performance. Reputation is comparable to our weighting of participants based on their adherence to formal criteria such as breadth in interest for topics, number of posts and ratings etc., and it determines the influence of participants on the discussion. Figure 2.3 features a comparison of projects and their reputation systems to our approach. Tausczik and Pennebaker have shown that building reputation is a very important incentive for online participation [80]. They have studied reasons why people contribute in communities of question-aswering websites using the community of MathOverflow, a site dedicated to research-level mathematics. In this context, of course, the level of expertise affects users' reported motivation to help others but it does not influence reputation building. Not only are individuals sensitive to the amount of recognition they receive, in addition recognition may consistently reinforce participation. Huberman and colleagues found a consistent pattern over time in which YouTube contributors posted more videos if two weeks earlier the videos they had uploaded had received more views [81]. In fact, recognition may create a positive feedback loop accounting for large discrepancies that are seen between users who contribute at high rates and those that contribute at low rates [82]. The significance of a properly designed reputation mechanism (karma) has been shown for Reddit [83], a social voting website where the incoming stream of links is voted up or down. A threat to all sites designed this way, however, is underprovision: when too many people rely on others to contribute without doing so themselves. 'karma' is built upon posting links that others like and vote for. Similarly, Slashdot [84], a website for sharing technology-related news, builds user reputation (or karma) through a number of activities, including moderation of comments, or posting comments that get high scores. The filtering of the content on Slashdot is facilitate by collecting ratings from selected readers who act as moderators assigning various labels such as "Informative", "Funny", or "Troll" to comments. Furthermore, Slashdot not only displays the score of each message, but also allows users to sort or filter the available messages based on those scores.

The aim of Slashdot and Reddit has been to nudge discussants to active participation. Nevertheless, Slashdot has not confirmed the relationship between high karma and the posting of top level comments or early postings. On the other side, these approaches have also revealed some issues in moderating content. User-based moderation with Slashdot has problems with overlooking or late detection of comments that are either very good or bad. Voting up on Reddit has led to overlooking half of the most popular links when they were first submitted. Hence, we can conclude that designing a proper reputation and rewarding system is very sensitive task which can heavily influence development of community and content quality. When designing our weighting scheme we have confined to formal criteria that should promote desirable behavior of participants such as breadth in discussing issues, number of original and focused comments, balanced tone and clear style of writing.

2.5 Comparative analysis of projects

In this section we give an overview of different projects that relates to our work (see Table 2.3). These different projects overlaps with our project in various aspects. For this purpose, we compare our approach to other approaches using different criteria such as whether they are argumentation based, if they employ different types of ratings, implement some reputation mechanisms or aim at decision-making.

Below: Platform Right:Parameters	*Type*	*Interface*	*Argumentation-based*	*Multi-facet ratings*	*Solutions*	*Reputation*	*Decision-making*
Digg	*news*	-	-	*no (dig/bury votes)*	-	*yes (the most influential voters)*	-
Reddit	*news, social networking*	-	-	*no (up/down votes)*	-	*yes ("karma", based on the votes of others)*	-
Epinions	*consumer review site*	-	-	*yes (different categories)*	-	*yes (usefulness of reviews - no control of abuse)*	-
Slashdot	*technology related news*	-	-	*yes (descriptors: "redundant", "troll", user tags)*	-	*yes ("karma" -votes of others)*	-
Deliberatorium	*discussion board*	-	*complex argument mapping*	*Rate quality of arguments/ideas*	*yes (collect ideas and arguments)*	*yes (based on ratings of authored questions and answers)*	*no*
ConsiderIt	*discussion board*	*simple, intuitive*	*yes (pro and contra arguments)*	*no*	*no*	*yes (based on the usage of arguments added, but no repetition control)*	*yes (calculates standing between the two opposite options)*
Deliberation forum model	*discussion board*	*based on conventional forum, simple, intuitive*	*yes (pro and contra arguments, proposals, proposal extensions)*	*yes (content, writing, tone)*	*yes (collect ideas, proposals for solution as well as arguments)*	*yes (based on formal criteria)*	*yes (based on comprehensive scoring model)*

Figure 2.3: Comparison of related project

2.6 Decision-making

A research problem of group decision-making is addressed in the literature. Decision-making is a process of arriving at a conclusion regarding a specific issue based on the opinions of multiple individuals. Consensus of the participants is an important indication of group agreement or reliability [85], [86]. Good measure of consensus should fully reflect the real behavior and opinion of the community. Traditional methods usually use numerical variance of participants opinion to calculate the consensus. That means verbal opinions are also transformed into numerical values and the variance indicates the disagreement degree. Nevertheless these method are often heavily criticized as not precise enough and misleading considering the interval scale assumed for ordinal verbal responses. On the other side, several works propose using the concept of entropy. In the work [87] authors propose a value-function approach for transforming verbal opinions into values on an interval scale

and measuring group consensus based on the value variability. To achieve consensus has appeared to be more difficult in computer-mediated discussion groups [88], [89],[90]. The reason is often that they need more time to make a decision and are more prone toward conflict.

Currently, there is an increasing number of computer-mediated social interactions related not only to decision-making. It has been shown that computer mediated discussion tend to be more focused on the task and less distracted by personal consideration [90] generating more ideas since group members can simultaneously participate. Murrel has explored the impacts of computer-based communication and group performance depending on the structure of communication systems [91]. Two synchronous systems for group decision-making with different immediacy of interactions and feedback have been examined. Even with these early discussion platforms, results shows that groups produce decisions superior to average initial individual solutions each user has submitted before joining the discussion. The comparison of results relies on a ground truth (expert opinions). In our settings there is no objective truth criterion, so the assessment is more complex and challenging.

Past research on computer-mediated group interactions indicates that the presence of a facilitator enhances the quality of group discussions [92]. Hilmer and Dennis [93] confirm that groupware increases the exchange of information for groups, but additional comments do not necessarily lead to better decisions in that context. The study explores groupware processes that require group members to categorize information. Different groupware processes have different effects on attention to and integration of information, and ultimately on decision quality. We argue that a moderation added to bring structure in participants discussion can improve the overall quality of discussion and facilitate more efficient decision-making. In our model we have foreseen some manual moderation when it comes to collecting proposals whereas the majority of the moderation activities is done by participants either by sorting out their posts or ratings other participants' posts.

2.7 Formal model and game-theoretic analysis

Models are frequently evaluated by their ability to estimate and observe phenomenon over a specified range. Obtaining valid inputs and validating outputs are critical steps in any modeling and simulation endeavor [94], [95]. There are no standard definition or procedure to validate models. Furthermore, building and validating social systems models is not a trivial task. Non-statistical models, especially agent-based and systems dynamic models have often been criticized for relying extensively on informal, subjective and qualitative validation procedures or no validation at all [96]. Gilbert on the other side claimed that "to validate a model completely, it is necessary to confirm that both the macro-level relationships are as expected and the micro-level behaviors are adequate representations of the actors' activity" [97]. Several macro or abstract level validation approaches have been proposed for social system models, including [98]: (1) theoretical verification or internal validation by subject matter expert determining conceptual validity [99]; (2) external validation against real world comparing the results from the model to observations in the real world [99]; and (3) cross-model validation that compares different models (e.g., [94]). Social system models, by definition, are complex, with imprecise, incomplete and inconsistent theories [100]. No

model is absolutely correct in the sense of a one-to-one correspondence between itself and real life, especially when it comes to agent-based and human behavior varieties. Hence, modeling should be treated as an iterative process bringing higher and higher understanding.

Game-theoretic models are used widely in social and behavioral sciences. Game-theory itself is consisted of different models, mathematical abstractions used to understand various interactions which can be observed as a game. A game in the general sense can refer to any competitive activity where players behave according to certain rules. Game theoretic reasoning pervades economic theory and is used widely in other social and behavioral sciences. The work presented by Skyrms in [101] is one of the first proposals of a game-theoretic model for deliberation. Considering the fact that game theory aims to help us understand situations in which decision-makers interact, the author describes deliberation as a dynamic process based on the concept of non-cooperative game. In group decision-making as a form of strategic situation, each player's optimal act is depending on the acts of others. Thus, his strategy is merely dependent of his expected utility and expectation about other participants' behavior. Another article by the same author [102] offers insights regarding the dynamics of the deliberation process and its steady point, the so-called deliberation equilibrium. A decision-maker is at a deliberational equilibrium if his state of indecision does not change under further deliberation. Deliberational dynamic is determined by two processes: (i) the way in which information generated by deliberation leads to modification of the expected utilities of the acts; and (ii) the dynamical law which modifies the state of indecision in response to the expected utilities. Skyrms and Bicchieri use the concept of deliberation based on expected utility [101], [102], [103].

Game-theoretic models are also used to explore the motivation to contribute and optimal designs of reward systems in various online social systems, e.g., forums, knowledge-sharing websites, and crowdsourcing projects, despite many recent innovations regarding such platforms. Singh et al. propose 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 [104]. 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. Ghosh and Hummel propose a ranking mechanism that maximizes the utility of the 'game owner' and incentivizes participants to give high-quality contributions [6]. 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 qualities 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 in the presence of strategic contributors. Being aware of this distinction, we target at valuation schemes that nudge discussants to contribute and give original arguments and honest feedback. Archak 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" [105].

Researhers, Yang et al. examine behavior of users on a Chinese web-based knowledge-sharing market, Taskcn.com [106]. 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. In their work DiPalantino et al. provide a game-theoretic model of multiple simultaneous crowdsourcing contests where agents select among, and subsequently compete in, several contests offering various rewards [107]. They model crowdsourcing contests as all-pay auctions with incomplete information on contestant skills.

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

3 Argumentation and Reasoning

Argumentation is vital for communication. The theory had its origin in foundationalism, a theory of justification or reasoning in the field of philosophy. Communication is very important for human beings. In fact, through communication the human beings begin to express their thoughts and thus it has played a significant role in our evolution.

In early times, the argumentation was based on oration and logic. One of the first to work on these theories was Aristotle. Aristotle's logic, especially his theory of the syllogism, has had a great influence on the history of modern thought. Although it did not always hold this position, in later antiquity, Aristotle's logic became dominant. Aristotle's theories have (in the opinion of many) a number of similarities of approach and interest with modern logic. Aristotle's logical works contain the earliest formal study of logic that we have. Kant claimed that nothing significant had been added to Aristotle's views in the span of two millennia. The rise of modern formal logic brought to light many serious limitations of Aristotle's logic. Nevertheless, scholars in modern formal techniques have come to view Aristotle with new respect, not so much for the correctness of his results as for the remarkable similarity in spirit between his work and modern logic [108]. The Aristotle's logic is built up around one principle: the deduction (*sullogismos*). Each of the "things supposed" is a premise (*protasis*) of the argument, and what "results of necessity" is the conclusion (*sumperasma*). These terms correspond to a modern notion of logical consequence: X results of necessity from Y and Z if it would be impossible for X to be false when Y and Z are true. This can serve as a general definition of "valid argument".

During 1960-1970 several scientists such as Perelman tried to develop the techniques used by people to get support of others for their views and opinions. That resulted in different approaches to argumentation. Perelman and Olbrechts-Tytecha presented a "new rhetoric" that reintroduced argumentation into rhetoric and reason theory of argumentation by establishing its link to Greek rhetoric and dialectic [46]. The traditional logic and reasoning alone cannot help resolve all disputes such as the consideration of some irrational elements in argumentation such as emotions. Restricting logic to the examination of the proofs termed 'analytical' by Aristotle, together with the reduction of dialectical proofs makes it inapplicable in real-life context. Hence, Perelman and Olbrechts-Tytecha claimed that argumentation is audience-centered, not form-centered and as such:

- has the goal of persuading an audience
- is more influenced by Aristotle *ethos*, authority and credibility of a rhetor
- dictates that the presumption and "burden of proof" are dictated by the audience, not the question or rhetor
- relies on the "community of minds", recognized by some authors as identification

- uses strategies associated with sophistry

Perelman and Olbrechts-Tytecha acknowledged that their theory of argumentation and with these proofs that fall outside of formal logic sound sophistic, would typically be dismissed as a "misleading form of reasoning". Some other authors also agreed on rejection of this strict formalism. Toulmin argues that logicians took analytical syllogisms as a paradigm and build their systems of formal logic completely on this foundation [48]. The truth is not a universal concept (thus the reason that the traditional syllogism will not work for philosophy and logic). Instead, each community has qualifiers or rebuttals for the idea of universal truth that make their own truth undefinable through formal syllogisms. Toulmin extended the traditional *premise-claim* model with additional elements: *warrant*, *backing*, *qualifier* and *rebuttal* (see Figure 3.1). *Claim or conclusion* is founded on the set of grounds (fact, evidence, data), whereas a statement that move a ground to a claim is *warrant* which in case of not being convincing enough can be supported by a *backing*. On the other hand, *rebuttal* are statements that recognize restriction applied to a claim. In the end, qualifiers are words or phrases expressing the degree of certainty of a claim such as "possible", "probably", "impossible", "certainly", "necessarily". Toulmin differs arguments that can be validated by their content and their form. His focus is on a different type of argument, called practical arguments or substantial arguments intended to focus on the justificatory function of argumentation, as opposed to the inferential function of theoretical arguments. While theoretical arguments make inferences based on a set of principles to arrive at a claim, practical arguments first find a claim of interest, and then provide justification for it.

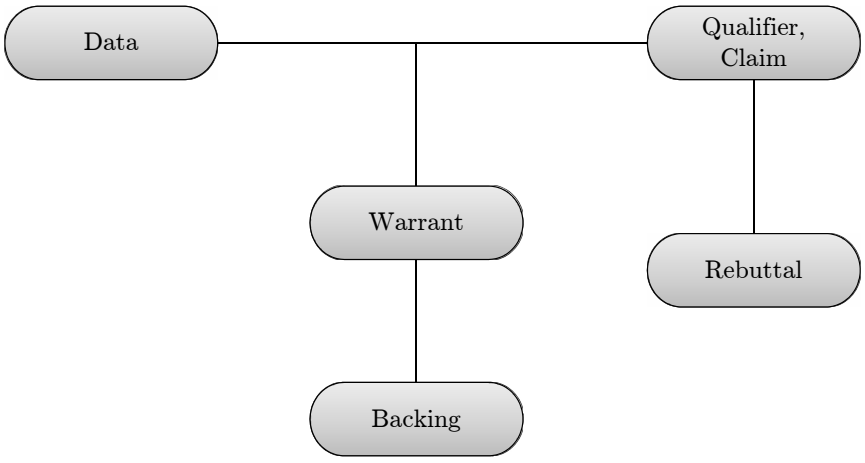


Figure 3.1: Toulmin's model of arguments

Based on these considerations we can distinguish *formal* and *informal* logic. *Formal* logic focuses on the premise-conclusion relation in exclusion to the question of premise acceptability by sorting them as: true (logical truths) and false ones ((logical falsehoods). Every other premise beyond these two ones cannot be analyzed with formal logic [109].

The trends in modern society are that the problems get more complex and the context is constantly changing. Dealing with such complex and important issues may require more advanced reasoning, critical thinking [110]. Informal logic tries to develop a logic that can assess and analyze the arguments that occur in natural language ("everyday", "ordinary language") discourse. Discussions can be various involving scientific, legal, and other technical forms of reasoning and different notions such as mathematical language, but the aim is a comprehensive approach that can explain and evaluate the arguments found in discussion, debate and disagreement as they appear in daily life - in social and political commentary; in news reports and editorials in the mass media (in newspapers, magazines, television, the World Wide Web, twitter, etc.); in advertising and corporate and governmental communications; and in personal exchange [111].

In following chapter we are going to give an overview of broad research field of informal logic. As a first step, we introduce base terms of formal logic that are also used in informal logic.

3.1 Basic terms in logic

Logic is non-empirical science of reasoning. It investigates the process of actual reasoning and its main task is to distinguish correct reasoning from the incorrect one. We can define reasoning as a mental activity called also *inferring* or making *inferences*. To infer means to draw a *conclusion* from a *premise*. *Premise* can be seen as a piece of information or data, or a fact. Output of the inferring process is a conclusion. The concern of logic is whether a conclusion if correct. Inferences are made on the basis of various sorts of things – called *statements*, which are noting else than declarative sentences in natural language.

Definition (Argument). An argument is a collection of statements, one of which is designated as the conclusion, and the remainder of which are designated as the premises.

Usually, premises of an argument are intended to support the conclusion of the argument. Argument is a method to increase the reliability in communication. Argumentation is also an activity of reason. When people argue, they use reasoning to assess and accept the conclusion. Communication is perfected with proper reasoning and also a true conclusion is supported by fair arguments.

Reasoning can be *deductive* or *inductive*. Thus, we distinguish *deductive logic* and *inductive logic* that differ by inferences. *Inductive logic* investigates the process of drawing probable though *fallible* conclusions from premises. *Inductive logic* relies on statistical or probabilistic reasoning. In *deductive logic*, the task is to distinguish deductively correct arguments from deductively incorrect arguments. An argument deductively correct, is certainly inductively correct, whereas inductively correct argument does not have to be also deductively correct. A *deductive argument* is based on premises that are meant to provide a guarantee of the truth of the conclusion. On the other hand, an *inductive argument* provides reasons that support the conclusion's probable truth. Non-deductive logic is reasoning using arguments which support the premises but not necessarily entail it. The standards for evaluating non-deductive arguments may rely on different or additional criteria than truth, for example, their persuasiveness. A *defeasible argument* is an argument which in the

context of certain counterreasons does not justifies the conclusion any more. Accordingly, *defeasible reasoning* is a kind of reasoning that is based on reasons that are *defeasible*, as opposed to the indefeasible reasons of deductive logic. *Defeasible reasoning* is a particular kind of non-demonstrative reasoning, where the reasoning does not produce a full, complete, or final demonstration of a claim, i.e., where *fallibility* and correctness of a conclusion are acknowledged.

Logic analyzes and classifies arguments according to their *form*, as opposed to their *content*. In that context, there are two important questions for reasoning:

- Question 1: Are all the premises true?
- Question 2: Does a conclusion follow from premise?

Definition (Factually Correct Argument). An argument is factually correct if and only if all of its premises are true.

Definition (Valid Argument). An argument is valid correct if and only if its conclusions follow from its premises.

Definition (Sound Argument). An argument is sound if and only if it is both factually correct and valid.

Related to evaluation of arguments by form and content, a *factually correct* arguments has *good content*, whereas a *valid* one has a *good form*. Hence, a *sound* argument has both *good form* and *good content*.

Fallacies are types of argument or expressions which are held to be of an invalid form of an argument or contain errors in reasoning. There is not yet any general theory of fallacy or strong agreement among researchers of their definition or potential for application but the term is broadly applicable as a label to certain examples of error, and/or various ambiguities.

In the end, arguments can be *formal* and *informal*. *Informal arguments* are the subject of study in i*nformal logic* and they are presented in everyday language and discourse. *Formal arguments* on the other side are studied in *formal (symbolic) logic* commonly referred as mathematical logic. *Formal arguments* are expressed in *formal language* and formal logic is based on *implication* and *inference*. While the connection between claims, premises, warrants, relations of implications and conclusion is more visible in formal arguments, informal logic relies on the study of argumentation that finds these elements.

3.2 Informal Logic

Informal logic attempts to develop a logic that can analyze and rate arguments expressed in natural language. Discussion may involve different types of reasoning, scientific, legal or other technical forms of reasoning but it is crucial to evaluate these arguments as they appear in real life, social or political topics, news and in the end in private life.

Informal logic combines traditional inference approach with the study of broad range of topics relevant for informal reasoning such as argument identification, "burden of proof", the

empirical study of argument; diagramming, cognitive bias; the role of emotion in argument etc. Hansen focuses on informal inference, whereas many involve multidisciplinary approach in order to understand and/or model natural language reasoning found in formal logic, cognitive psychology, rhetoric, dialectics, computational modeling, and a range of other fields [109]. The interdisciplinary study of informal reasoning is the basis of so called *argumentation theory*.

In early formulations, informal logic was sometimes understood as a theoretical alternative to formal logic debating whether informal logic fit within the study of "real" logic. Nowadays, the field of informal logic enjoys a more conciliatory relationship with formal logic. It attempts to understand informal reasoning expressed in natural language. Research may employ formal methods and the question whether the accounts of argument which characterize informal logic can in principle be formalized is a source of active investigation. Some recent works in computational modeling attempt to implement informal logic models of natural-language reasoning. They suggest that defeasible (non-monotonic) logic, probability theory, and other non-classical formal frameworks may be suited to this task.

3.2.1 History

Informal logic has a number of historical precedents. The origins of informal logic are found in North America in the 1970s. It is often said that informal logic was a consequence of rising political and social movements in the society in 1960-ties. The practical purpose of fostering logic in everyday life was augmented by the social commentary and debates in newspapers, the mass media, advertisements and political campaigns.

In North America and other English speaking countries, the "Critical Thinking Movement", aimed to inform and improve public reasoning and debate by promoting models of education which emphasize the critical examination of beliefs and decisions, and the development of the skills that this requires.

Though the focus on real life examples was anticipated in Hamblin's Fallacies (1970) and Toulmin's Use of Arguments (1958), informal logic started with the work of Johnson and Blair at the University of Windsor. In the textbook, Logical Self-Defense (1977) gave some insights into the logic of informal reasoning. They have established a journal (now Informal Logic) to introduce a new discipline within the field of theoretical discussion. In forty years, a large body of literature was collected with evolving set of topics and issues defining a field. The field has become broad and involves issues such as fallacies, argument schemes; the rhetorical features of argument; dialectical obligations; dialogue theory; kinds of arguments (deductive, inductive, conductive); the role of images and diagrams in arguments; empirical studies of argument; communication in argumentative contexts; and the history of argument analysis. Today, the continued development of informal logic increasingly incorporates approaches to discourse and argumentation found in cognate disciplines and fields like Speech Communication, Rhetoric, Linguistics, Artificial Intelligence, Cognitive Psychology, and Computational Modeling.

3.2.2 Informal argument

Similarly as in classical logic, argument in informal logic has been understood as an attempt to present evidence for a conclusion. It does so by providing premises ("propositions" or claims or some sort) that support the conclusion. Hitchcock provides a precise account of this conception, defining an argument as "a claim-reason complex" consisting of:

(i) an act of concluding,

(ii) one or more acts of premising (which assert propositions in favor of the conclusion), and

(iii) a stated or implicit inference word that indicates that the conclusion follows from the premises [112].

Hitchcock's account of argument is broad. It allows premises and conclusions to be any speech acts which assert the truth of a proposition (including acts like suggesting, hypothesizing, boasting, and deducing), and recognizes that arguments in natural language frequently occur without an explicit inference indicator like "since" or "therefore". The scope of informal argument is significantly widened by recognizing visual (and other kinds of non-verbal) arguments. The motivation is to have some theoretical means for understanding and assessing informal arguments. For the same reason, many informal logicians now distinguish between two senses of 'argument' which are commonly designated: "*argument-1*" and "*argument-2*". *Argument-1* is argument in the traditional premise and conclusion sense. *Argument-2* is argument understood as the disagreement and/or exchange in which argument-1 is typically embedded. Sometimes the difference between these two kinds of argument is expressed by describing argument-2 as process or transaction, and argument-1 as the product that results from it. Informal logic has paid increasing attention to argument-2 as discussions in the field have evolved, for the simple reason that the assessment and analysis of argument-1 often requires an understanding of argument in this broader sense. In judging the reasonableness of a particular example of argument-1, and the extent to which it is appropriate or convincing, we must frequently consider the argument-2 that gives rise to it.

One impetus for the development of informal logic has been the view that natural language arguments do not fit the deductive framework emphasized in traditional logic. The extent to which informal arguments can be understood as deductive arguments has, therefore, been a source of significant debate within informal logic. "Natural Language Deductivism" (NLD) is the view that all informal arguments should be interpreted as attempts to create deductively valid inferences. If the premises of a deductively valid argument are true, its conclusion must be true (i.e., cannot be false). Deductive arguments have traditionally been associated with logical and mathematical reasoning thought to produce certain or necessary conclusions, but good deductive arguments in informal contexts typically yield conclusions that are reasonable or plausible because they rely on premises which are reasonable or plausible (rather than certain).

The goal of natural language deductivism is an approach to informal arguments which allows one to effectively and efficiently assess the support they provide for their conclusions. It suggests that one should do so by:

(i) interpreting an argument as an attempt to construct a deductively valid inference; and then

(ii) assessing the credibility of the premises of the argument

3.2.3 Fallacy theory and argumentation schemes

Early work in informal logic favored fallacies as a way of assessing informal arguments. Walton claims that "many of the fallacies are misuses of presumptive inference" [113]. A fallacy is a pattern of poor reasoning which appears to be (and in this sense mimics) a pattern of good reasoning [109]. Theoretical discussions of fallacies have not produced an agreed-upon taxonomy, but there is a common set of fallacies which are typically used in the analysis of informal arguments. They include formal fallacies like affirming the consequent and denying the antecedent; and informal fallacies like *ad hominem* ("against the person" used to attack someone's argument by raising questions about that person's character or personal situation), *ad bacculum* ("appeal to force" prevents critical discussion by closing off free expression of opinion), "hasty generalization", and "two wrongs" (as in "two wrongs do not make a right"). In the research literature, Woods and Walton have discussed the definition, analysis and assessment of a variety of fallacies in a series of articles and books (i.e. [114]). Van Eemeren and Grootendorst have proposed a "pragma-dialectical" theory which analyses fallacies as violations of the rules of critical discussion (discussion which aims to critically resolve a difference of opinion)[115].

Non-fallacy approaches to informal argument make informal logic comparable to classical formal logic. In both cases, it is very important to identify general criteria for good reasoning and argument schemes that incorporate specific forms of reasoning. Argumentation schemes focus on validity and soundness, and deductive argument schemes encapsulate rules of inference like *modus ponens* ("Affirming the Antecedent"), double negation, *modus tollens* ("Denying the Consequent"), etc. In the case of informal logic, the standard criteria for good argument, as already mentioned in the context of formal argument, can be reduced to: (i) premise acceptability and (ii) a conclusion that follows from the premises. The second criterion is typically understood in terms of relevance and sufficiency, making a good argument an argument with premises that are relevant to the conclusion and sufficient to establish it as acceptable. Within informal logic, the key argument schemes discussed include arguments from authority, causal reasoning, arguments by analogy, and various forms of moral argument.

Argumentation schemes are forms of inference from premises to a conclusion for kind of arguments used in everyday conversational exchanges in which one party is trying to get another to come to accept a conclusion that is at issue. They represent patterns of deductive and inductive reasoning in some instances, but typically they represent defeasible inferences of a kind that are useful heuristics for moving to a plausible hypothesis under conditions of uncertainty and lack of knowledge [116]. Prakken considers argument schemes are essentially logical constructs, which evaluate arguments primarily based on the form of a logic. More specifically, most argument schemes are defeasible inference rules and their critical questions are pointers to counterarguments, so that the logic governing the use of argument schemes should be a logic for non-monotonic, or defeasible reasoning

[117]. Schemes identify arguments, find missing premises, analyze arguments, and finally evaluate them. The tool used for evaluation is the set of appropriate critical questions matching each scheme [47]. Walton associates a list of critical questions for schemes, pointing out of the informal fallacy or fallacies that typically follow that scheme [113]. He sees presumptive reasoning in general as a kind of reasoning from ignorance, for we are forced to use presumptive reasoning when we lack complete knowledge. Different forms of presumptive reasoning can be captured in argumentation schemes, and how lists of critical questions may be drawn up for each scheme to identify the qualifications that must be met for such reasoning to be cogent (or at least plausible). If the qualifications identified by the critical questions of a given argumentation scheme are ignored, and the reasoner or arguer goes ahead and draws or urges the conclusion when those qualifying circumstances are not present, then the argument will be fallacious. Since many fallacies consist of violations of the conditions of argument schemes, those fallacies can all be seen as special cases of the generic fallacy of ignoring qualifications.

Although argumentation schemes have shown to have limited application in public deliberation, the research field of informal logic is very useful to outline common problems of evaluating arguments expressed in "natural language" i.e. fallacies as different kind of errors in reasoning, power of authority or persuasiveness, emotions, lack of knowledge or certainty of truth and other difficulties that were completely neglected in formal logic. In the end, the main goal of the informal logic is to support and encourage critical thinking even if it does not give perfect solutions, but rather provide recommendations how to improve everyday discourse.

4 Game Theory

Game theory is a branch of applied mathematics and economics that studies strategic decision-making. Game theory is "the study of mathematical models of conflict and cooperation between intelligent and rational decision-makers" [118]. Strategic nature comes from various parties, stakeholders involved, who have their own interest, goals and whose actions can affect one another and, thus, are interdependent. This interdependence causes each player to consider the other player's possible decisions, or strategies, and select his own strategy. A solution to a game describes the optimal decisions of the players, who may have similar, opposed, or mixed interests, and the outcomes depend on these decisions. An alternative term for the game theory is interactive decision theory [119]. Hence, it attempts to determine mathematically and logically the actions that "players" should take to secure the best outcomes for themselves in a wide array of "games" [120].

Game theory studies various games such as card games, chess etc., but it has also been applied to complex business issues or military strategy. Its main applications are in economics, political science, and psychology, as well as in logic, computer science, and biology. All these games share the common feature of interdependence. That is, the outcome for each participant depends on the choices (strategies) of all other participants. A game in the everyday sense is defined a "a competitive activity in which players contend with each other according to a set of rules". Thus, the scope of game theory is vastly larger and can be applied to many situations: firms competing for business, elections with political opponents competing, bidders in an auction.

In general, the value of game theory lies in understanding the interactions and likely outcomes when the end result is dependent on the actions of others who have potentially conflicting motives. Game theory allows structured analysis of complex multiple players issues including the identification of the best attainable outcome, threats and promises available to different players and the prediction of the likelihood of actions and reactions of other players. In other words, the concepts of game theory provide a language to formulate, structure, analyze, and understand strategic scenarios [121].

Game theory is divided in two branches, the *non-cooperative* and *cooperative games*. The difference between these two branches is based mainly on the way how interdependence among players is formalized. The non-cooperative theory might be better described as procedural, whereas the cooperative theory as combinatorial interdependence. This indicate the main distinction, first defines various actions that are available to players while the second describes the outcomes that result when the players come together in different combinations.

A cooperative game is a game where groups of players ("coalitions") may enforce cooperative behavior, and the game is a competition between coalitions of players, rather than between individual players. The idea behind cooperative games is to focus on the possibilities of

agreement. Thus, they are treated separately, the formalization is built with negotiation and enforcement procedures in the model and the results depends highly on the procedures, on the order of making offers and counter-offers.

On the other hand, non-cooperative games are the ones in which players make decision independently. Non-cooperative game theory is concerned with the analysis of strategic choices. The paradigm of non-cooperative game theory is that the details of the ordering and timing of players choices are crucial to determining the outcome of a game. Models of non-cooperative games represent players as self-interested agents who make choices out of their own interest. Cooperation in non-cooperative games is not necessarily out of question, and if players find it in their best interest the cooperation can also happen. Furthermore, the theory of non-cooperative games does not apply exclusively to situations in which the interests of different agents conflict as well as in non-cooperative games players interest align.

Rationality is a basic assumption of game theory and all (pure) game-theoretic models are based on it. A rational player chooses an action which gives the outcome he most prefers given his expectations about the behavior of other players.

The consistency and mathematical foundations of game theory make it a sophisticated tool for modeling and designing automated decision-making processes in interactive environments. The automation of strategic choices enhances the need for these choices to be made efficiently, and to be robust against abuse. Game theory addresses these requirements. As a mathematical tool for a decision-maker a strength of game theory is the methodology it provides for structuring and analyzing problems of strategic choice. The process of formal modeling a situation as a game requires a decision-maker to enumerate explicitly the players and their strategic options, and to consider their preferences and reactions. The discipline involved in constructing such a model already has the potential of providing the decision-maker with a clearer and broader view of the situation. This is a "prescriptive" application of game theory, with the goal of improved strategic decision-making [121].

The understanding, that game theoretic models give, is particularly relevant in the social, political, and economic fields. Studying game theoretic models may also suggest ways to improve our welfare. These models can be verbally described but turning them into mathematical forms, they improve the conciseness and precision of our consideration.

4.1 History of game theory

The first example of a formal game-theoretic analysis is the study of a duopoly by Antoine Cournot in 1838 and the mathematician Emile Borel was the first to suggest a formal theory of games in 1921. The work was continued by the Princeton mathematician John von Neumann in 1928 as a "theory of parlor games". Game theory was established as a field after publication of "Theory of Games and Economic Behavior" by von Neumann and the economist Oskar Morgenstern in 1944. In the book these two authors provided the basic terminology and the problem setup that is still nowadays in use. The second edition of this book introduced an axiomatic theory of expected utility, which allowed mathematical statisticians and economists to treat decision-making under uncertainty. Modern game

theory began with an idea of the existence of mixed-strategy equilibria in two-person zero-sum games and its proof by John von Neumann.

Later, in 1950, John Nash demonstrated that finite games always have an equilibrium point, at which all players choose actions which are best for them given their opponents' choices. This central concept of non-cooperative game theory has been a central point of analysis since then. In the 1950s and 1960s, game theory was broadened theoretically and applied to problems of war and politics. Since the 1970s, it has driven a revolution in economic theory. Additionally, it has found applications in sociology and psychology, and established links with evolution and biology. Game theory received special attention in 1994 with the awarding of the Nobel prize in economics to John F. Nash, John C. Harsanyi, and Reinhard Selten for their pioneering analysis of concept of equilibria in non-cooperative games.

At the end of the 1990s, a high-profile application of game theory has been the design of auctions. Prominent game theorists have been involved in the design of auctions for allocating rights to the use of bands of the electromagnetic spectrum to the mobile telecommunications industry. These applications turned out to be not only efficient comparing to traditional practices, but additionally have expanded in the United States and Europe [121].

4.2 Definition of games

A game is a formal model of an interactive situation. Typically, it involves several players, whereas a game with only one player is referred as a decision problem [121]. Players in the game are self-interested agents. A decision maker chooses the best actions according to his preference among all the available ones. Actions are set of moves that under some circumstances are available to the decision-maker, and a specification of the decision-makers preferences. At a given point, a player has at least a subset of these actions available and he evaluates them according to his preferences [122].

A dominant approach to model an agent's interests is the utility theory. The underlying theoretical approach of the utility theory aims to quantify an agent's degree of preference across a set of available alternatives. An important goal of this concept is to understand how these preferences change when an agent faces uncertainty about the behavior of other agents. A utility function transforms the real world outcomes to numbers. These numbers measure agent's utility in the given states, e. g. the higher is the number, the higher would be agents' utility. When the agent is uncertain about which state of the world he faces, his utility is defined as the expected value of his utility with respect to the appropriate probability distribution over states. Utility theorists respond to such questions by showing that the idea of utility preferences is based on a more basic concept of preferences. The most influential theory is the one from c, and thus the utility functions are sometimes called von NeumannMorgenstern utility functions to distinguish them from other varieties. In economic theory a payoff function that represents a consumer's preferences is often referred to as a "utility function". A decision-maker's preferences can be represented by many different payoff functions.

Let O denote a finite set of outcomes. For any pair of outcome o_1, $o_2 \in O$, let $o_1 \succeq o_2$ represent the agent's weakly preference for o_1 to o_2. Next, $o_1 \sim o_2$ denote the proposition

that the agent is indifferent between o_1 and o_2. Finally, by $o_1 \succ o_2$, denote the agent's strictly preference for o_1 to o_2. Preferences interact with uncertainty about the selection of the outcome based on a lottery principle. Formally, lottery principle is a probability distribution over outcomes $[p_1 : o_1, \dots p_k : o_k]$, where each $o_i \in O$, each $p_i \geq 0$ and $\sum_{i=1}^{k} p_i = 1$. According to [123] the axioms of utility theory are following:

Axiom (Completeness) introduces the ordering between possible outcomes. For every pair of outcomes, a player prefers one or the other or he is indifferent between them.

$\forall o_1, o_2, o_1 \succ o_2$ or $o_1 \succ o_2$ or $o_1 \sim o_2$

Axiom (Transitivity) holds the concept of rationality of players who have transitive preferences.

If $o_1 \succeq o_2$ and $o_2 \succeq o_3$, then $o_1 \succeq o_3$.

Axiom (Substitutability) represents a preposition that if an agent is indifferent between two outcomes, he is also indifferent between two lotteries. Outcomes perceived similarly by agent and with similar or the same probability are mutually substituted.

If $o_1 \sim o_2$, then for all sequences of one or more outcomes $o_3, \dots, o_k$ and sets of probabilities p, $p_3, \dots, p_k$ for which $p + \sum_{i=3}^{k} p_i = 1$, $[p : o_1, p_3 : o_3, \dots, p_k : o_k] \sim [p : o_2, p_3 : o_3, ..., p_k : o_k]$.

Let $P_l(o_i)$ denotes the probability of outcome o_i in lottery l. For example, if Lottery l is the composed lottery $[0.3 : o_1; 0.7 : [0.8 : o_2; 0.2 : o_1]]$, then $P_l(o_1) = 0.44$ and $P_l(o_3) = 0$.

Axiom (Decomposability) states that an agent is always indifferent between lotteries that induce the same probabilities over outcomes.

If $\forall o_i \in O, P_{l_1}(o_i) = P_{l_2}(o_i)$ then $l_1 \sim l_2$.

Axiom (Monotonicity) claims that agents prefer more of a good thing. When an agent prefers outcome o_1 to o_2 and considers two lotteries, he also prefers the one that gives the higher probability to the more preferable outcome.

If $o_1 \succ o_2$ and $p>q$ then $[p : o_1, 1 - p : o_2] \succ [q : o_1, 1 - q : o_2]$.

Axiom (Continuity)

If $o_1 \succ o_2$ and $o_2 \succ o_3$, then $\exists p \in [0, 1]$ such that $o_2 \sim [p : o_1, 1 - p : o_3]$.

Theorem (von Neumann and Morgenstern, 1944). If a preference relation satisfies the axioms completeness, transitivity, substitutability, decomposability, monotonicity, and continuity, then there exists a function $u : O \to [0, 1]$ with the properties that:

1. $u(o_1) \geq u(o_2)$ iff $o_1 \succeq o_2$, and
2. $u^*([p_1 : o_1, ..., p_k : o_k]) = \sum_{i=1}^{k} p_i u(o_i)$.

The statement (1) follows trivially (both sides of the implication are always true), and statement (2) derives from decomposability. Otherwise, there must be a set of one or more most-preferred outcomes and a disjoint set of one or more least-preferred outcomes. When these axioms are satisfied utility function must exist for every set of preferences. The range of utility function is not important and any positive transformation yields another utility function for the same agent [123].

The above presented formalism is based on the utility function approach as a measure of agents' preferences. Another approach to model a player's behavior is the theory of rational choice [122]. The theory is an important component of many models in game theory and according to it a player's decision about action does not depend on any qualitative characteristic of preferences. The theory of rational choice is enormously successful; it is a component of countless models that enhance our understanding of social phenomena.

4.3 Normal-form games

Games in game theory can be presented in different ways. The most common representation is normal form or strategic form. Strategic form introduces strategic interactions by presenting every player's utility for every state of the world, where states depend only on players' combined actions, strategies. In static games of complete, perfect information, a normal-form representation of a game is a specification of players' strategy spaces and payoff functions. A strategy space for a player is the set of all strategies available to that player, whereas a ***strategy*** is *a complete plan of action for every stage of the game.* The presented formalism of normal-form games is presented in a book by Shoham et al. [123].

Definition (Normal-form game) A (finite, n-person) normal-form game is a tuple (N, A, u), where:

- N is a finite set of n players, indexed by i;
- $A = A_1 \times \cdots \times A_n$, where A_i action is a finite set of actions available to player i. A normal-form game considers set of strategies for all players which fully specifies all actions in the game.

Each vector $a = (a_1, ..., a_n) \in A$ is called an action profile; $u = (u_1, ..., u_n)$ where $u_i : A \to R$ is a real-valued utility (or payoff) function for player i. Note that we have previously introduced utility functions to map the set of outcomes, not the set of actions. The implicit assumption is $O = A$.

A natural way to represent games is via an n-dimensional matrix. Namely, the strategic form (or normal form) is a way of describing a game using a matrix (see Table 4.1). The game is defined by exhibiting on each side of the matrix the different players (here players 1 and 2), each strategy or choice they can make (here strategies A and B) and sets of payoffs (values) they will each receive for a given strategy $(v_{1A}, v_{2A}; v_{1A}, v_{2B}; v_{1B}, v_{2A}; v_{1B}, v_{2B})$.

		Player 2	
		Strategy A	Strategy B
Player 1	Strategy A	v_{1A}, v_{2A}	v_{1A}, v_{2B}
	Strategy B	v_{1B}, v_{2A}	v_{1B}, v_{2B}

Table 4.1: Players' strategies

The strategic form is usually used for simultaneous games, where both players choose their strategies simultaneously. Simultaneous games imply there is a finite and imperfect information, and the rules of the game as well as each players payoffs are common knowledge. A well-known example of a simultaneous game described using the strategic form is the prisoners dilemma (see Figure 4.2). Two suspects are held in separate cells. Since there is enough evidence to convict each of them of a minor offense, but not enough evidence to convict either of them of the major crime unless one of them acts as an informer against the other (finks). If they both stay quiet, each will be convicted of the minor offense and spend one year in prison. If one and only one of them finks, she will be freed and used as a witness against the other, who will spend four years in prison. If they both fink, each will spend three years in prison. This situation may be modeled as a strategic game:

- Players The two suspects.
- Actions Each players set of actions is $\{Quiet, Fink\}$.[122]

		Suspect 2	
		Quiet	Fink
Suspect 1	Quiet	$(2, 2)$	$(0, 3)$
	Fink	$(3, 0)$	$(1, 1)$

Table 4.2: Prisoners' dilemma

The Prisoners Dilemma models a situation in which there are gains from cooperation (each player prefers that both players choose Quiet than they both choose Fink) but each player has an incentive to "free ride" (choose Fink) whatever the other player does. The game is important not because we are interested in understanding the incentives for prisoners to confess, but because many other situations have similar structures.

So far we defined the actions available to each player in a game, but not yet his set of strategies or his available choices. Certainly one kind of strategy is to choose a single action and this is called pure strategy. A choice of pure strategy for each agent is a pure strategy profile. Another, less obvious type of strategy: randomize the choice over the set of available actions according to some probability distribution [123]. Such a strategy is called a mixed strategy.

Definition (Mixed strategy). Let (N, A, u) be a normal-form game, and for any set X let $\prod(X)$ be the set of all probability distributions over X. Then the set of mixed strategies for player i is $S_i = \prod(A_i)$.

Definition (Mixed-strategy profile). The set of mixed-strategy profiles is the Cartesian product of the individual mixed-strategy sets $S_1 \times \ldots S_n$.

Definition (Expected utility of a mixed strategy). Given a normal-form game (N, A, u), the expected utility u_i for player i of the mixed-strategy profile $s = (s_1, ..., s_n)$ is defined as:

$$u_i(s) = \sum_{a \in A} u_i(a) \prod s_j(a_j).$$

4.3.1 Nash equilibrium

The term *Nash equilibrium* was first time introduced in a PhD thesis of John F. Nash Jr. The notion vastly expanded the scope of game theory, previously focused on two-player "strictly competitive" games (in which the players' interests are directly opposed) to a new class of strategic games.

The question that rises up in strategic games is which strategy will a player choose. We assume that players are rational and each player takes the best available actions. The selection of the best action merely depends on other players' moves. So the choice of an action relies on a player's beliefs about other players' behavior. Underlying assumption is that beliefs are based on his experience.

A Nash equilibrium is an action profile a^* with the property that no player i can do better by choosing an action different from a_i^*, given that every other player j adheres to a_j^*.[122]

Ideally the players in any given play of the game are drawn randomly, and a Nash equilibrium corresponds to a steady state. If, whenever the game is played, the action profile is the same Nash equilibrium a^*, then no player has a reason to choose any action different from the component of a^*; there is no pressure on the action profile to change. Hence, a Nash equilibrium incorporates a principle of a stable "social norm": if everyone else adheres to it, no individual would deviate from it [122].

Let a be an action profile, in which the action of each player i is a_i. Let a_i' be any action of player i (either equal to a_i, or different from it). Then (a_i', a_{-i}) denotes the action profile in which every player j except i chooses his action a_j as specified by a, whereas player i chooses a_i'. (The $-i$ subscript on a stands for "except i".) That is, (a_i', a_{-i}) is the action profile in which all the players other than i adhere to a while i "deviates" to a_i'. (If $a_i' = a_i$ then of course $(a_i', a_{-i}) = (a_i, a_{-i}) = a$). If there are three players, for example, then (a_2', a_{-2}) is the action profile in which players 1 and 3 adhere to a (player 1 chooses a_1, player 3 chooses a_3) and player 2 deviates to a_2'. Thus, the condition for an action profile a^* to be a Nash equilibrium is following: no player i has any action a_i for which he prefers (a_i, a_{-i}^*) to a^*. Analogously, for every player i and every action a_i of player i, the action profile a^* is at least as good for player i as the action profile (a_i, a_{-i}^*).

Definition (Nash equilibrium of strategic game with ordinal preferences). The action profile a^* in a strategic game with ordinal preferences is a Nash equilibrium if, for every player i and every action a_i of player i, a^* is at least as good according to player i's preferences as the action profile (a_i, a_{-i}^*) in which player i chooses a_i while every other player j chooses a_j^*. Equivalently, for every player i, $u_i(a^*) \geq u_i(a_i, a_{-i}^*)$ for every action a_i of player i, where u_i is a payoff function that represents player i's preferences.

A strategic game does not necessarily have a Nash equilibrium in case of pure strategies. In games with mixed strategies Nash equilibrium always exists. Furthermore, there can be many Nash equilibria. The definition of a Nash equilibrium is designed to model a steady

state among experienced players. An alternative approach to understanding players' actions in strategic games assumes that the players know each others' preferences, and considers what each player can deduce about the other players' actions from their rationality and their knowledge of each other's rationality.

The theory of Nash equilibrium, has two components:

1. the players act in accordance with the theory of rational choice, given their beliefs about the other players' actions,

2. these beliefs are correct.

The Prisoners Dilemma has attracted a great deal of attention by economists, psychologists, sociologists, and biologists. A huge number of experiments have been conducted with the aim of discovering how people behave when playing a game. To illustrate the Nash equilibrium we use the Prisoners Dilemma.

Please refer to the Table 4.2. The action pair *(Fink, Fink)* is a Nash equilibrium because (i) given that player 2 chooses *Fink*, player 1 is better off choosing *Fink* than *Quiet* (looking at the right column of the table we see that *Fink* yields player 1 a payoff of 1 whereas *Quiet* yields her a payoff of 0), and (ii) given that player 1 chooses *Fink*, player 2 is better off choosing *Fink* than *Quiet* (looking at the bottom row of the table we see that *Fink* yields player 2 a payoff of 1 whereas *Quiet* yields her a payoff of 0). No other action profile is a Nash equilibrium.

Introduced concept of Nash equilibria is based on a principle that any deviation by a player leads to an outcome worse for that player than the equilibrium outcome. The definition of Nash equilibrium, however, requires only that the outcome of a deviation be no better for the deviant than the equilibrium outcome. For a general game, an equilibrium is strict if each player's equilibrium action is better than all his other actions, given the other players' actions. Precisely, an action profile a^* is a strict Nash equilibrium if for every player i we have $u_i(a^*) > u_i(a_i, a^*_{-i})$ for every action $a_i \neq a^*_i$ of player i.

Beside the concept of equilibrium strategy which represents the steady state of the game, we introduce also the concept of *dominant strategy*. In simple words, a player's action "strictly dominates" another action if it is superior, no matter what the other players do.

Definition (Strict domination). In a strategic game with ordinal preferences, player i's action a''_i strictly dominates his action a'_i if $u_i(a''_i, a_{-i}) > u_i(a'_i, a_{-i})$ for every list a_{-i} of the other players' actions, where u_i is a payoff function that represents player i's preferences. In the Prisoners Dilemma, for example, the action *Fink* strictly dominates the action *Quiet*: regardless of any opponents action, a player prefers the outcome *Fink*.

If an action strictly dominates the action a_i, we say that a_i is strictly dominated. A strictly dominated action is not a best response to any actions of the other players: whatever the other players do, some other action is better. Since a player's Nash equilibrium action is a best response to the other players' Nash equilibrium actions, a strictly dominated action is not used in any Nash equilibrium. Thus, when looking for the Nash equilibria of a game, we can thus eliminate from consideration all strictly dominated actions.

Definition (Weak domination). In a strategic game with ordinal preferences, player i's action a_i'' weakly dominates an action a_i' if and $u_i(a''i, a_{-i}) \geq u_i(a'i, a_{-i})$ for every list a_{-i} of the other players' actions $u_i(a_i'', a_{-i}) > u_i(a_i', a_{-i})$ for some list a_{-i} of the other players' actions, where u_i is a payoff function that represents player i's preferences.

In a strict Nash equilibrium no player's equilibrium action is weakly dominated: every non-equilibrium action for a player yields a payoff less than does the equilibrium action, and, thus, does not weakly dominate the equilibrium action.

If every subject understands the game and faces incentives that correspond to the preferences of the player of a certain role, then a divergence between the observed outcome and a Nash equilibrium can be blamed on a failure of one or both of these two components or his beliefs about the other subjects' behavior. Experimental evidences so far have shown that there are games for which the theory works well and, and some for which it does not work that good. There are several issues that do play role in these experimental results:

1. Ensure to induce the preferences that are the subject of the study. In other words, each subject's preferences are matching with the role he is taking in the game. The standard way of inducing the appropriate preferences is to introduce monetary payoff as indicator of preferences. Assumption that player are self-interested rational agents might not be reasonable in some situation. Thus Nash equilibrium might not be appropriate for every study of strategic behavior.
2. Assessing an appropriate ratio between gain and effort might be also a critical point. I.e. monetary payments may not be necessary: under some circumstances, in a highly competitive culture such as USA it may be sufficiently motivating to obtain a high score, even if that score does not translate into a monetary payoff.
3. Finding a steady state of the game, in some cases can be not of a high relevance when the focus of the study is different i.e. the learning process of agents or effects of experienced players on their behavior and choice of the strategy.

Finally, as with any other theory, the theory of Nash equilibrium might appear as not fully correct.

4.4 Extensive games

The model of a strategic game reflects the sequential structure of decision-making. Decision-makers move sequentially and choose a plan of action once and for all; no matter how events unfold. The model of an extensive game, by contrast, describes the sequential structure of decision-making explicitly, and each decision-maker is free to change his mind as events unfold [122]. In other words, extensive games are dynamic games where players choose their actions sequentially. In case of *perfect information*, each player can *perfectly observe past actions*. Hence, extensive form games need not only to specify the set of players and their preferences but also the sequence of moves or actions at each stage of the game. Each possible sequence is a terminal history. As for a strategic game, ***player's preferences*** can be specified by giving *a payoff function* that represents them. In some situations an outcome is associated with each terminal history, and *the players preferences are naturally*

defined over these outcomes, rather than directly over the terminal histories. In the general definition, *outcomes are conveniently identified with terminal histories and preferences are defined directly over these histories*, avoiding the need for an additional element in the specification of the game.

An extensive game has four components:

- Players
- Terminal histories
- Player function
- Preferences for the players

The description of the game specifies the set of terminal histories and the player function, from which we can deduce the available sets of actions. The terminal histories of a game are specified as a set of sequences and not every set of sequences is a legitimate set of terminal histories.

Definition (Extensive game with perfect information). An extensive game with perfect information consists of:

- a set of players
- a set of sequences (terminal histories) with the property that no sequence is a proper subhistory of any other sequence
- a function (the player function) that assigns a player to every sequence that is a proper subhistory of some terminal history
- for each player, preferences over the set of terminal histories. The set of terminal histories is the set of all sequences of actions that may occur.

The set of terminal history is the set of all possible actions sequence that a player can undertake assigned by the player function.

Example (Entry game). An incumbent faces the possibility of entry by a challenger. (The challenger may, for example, be a firm considering entry into an industry currently occupied by a monopolist, a politician competing for the leadership of a party etc..) The challenger may enter or not. If it enters, the incumbent may either acquiesce or fight. See Figure 4.1.

The situation can be modeled as an extensive game with perfect information in which the terminal histories are *(In, Acquiesce)*, *(In, Fight)*, and *Out*, and the player function assigns the challenger to the start of the game and the incumbent to the history. In the situation described above, suppose that the best outcome for the challenger is that it enters and the incumbent acquiesces, and the worst outcome is that it enters and the incumbent fights, whereas the best outcome for the incumbent is that the challenger stays out, and the worst outcome is that it enters and there is a fight. Then the situation may be modeled as the following extensive game with perfect information.

- *Players.* The challenger and the incumbent.

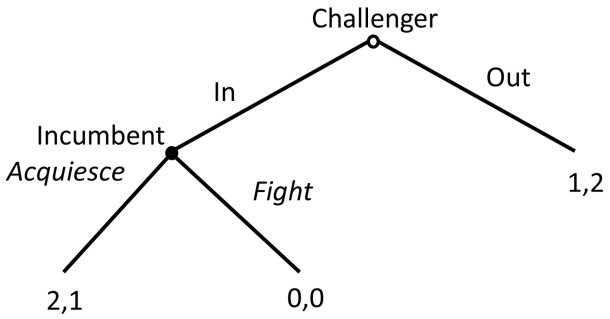


Figure 4.1: Extensive Form - Entry Game

- Terminal histories *(In, Acquiesce)*, *(In, Fight)*, and *Out*. Player function $P(\oslash) = Challenger$ and $P(In) = Incumbent$.
- *Preferences.* The challenger's preferences are represented by the payoff function u_1 for which $u_1(In, Acquiesce) = 2$, $u_1(Out) = 1$, and $u_1(In, Fight) = 0$, and the incumbent's preferences are represented by the payoff function u_2 for which $u_2(Out) = 2$, $u_2(In, Acquiesce) = 1$, and $u_2(In, Fight) = 0$.

4.4.1 Games with payoff

In the game in the example, available histories are $\oslash$, *In*, *Out*, *(In, Acquiesce)*, and *(In, Fight)*. Thus, the set of actions available to the player who moves at the start of the game, namely the challenger, is $A(\oslash) = In, Out$, and the set of actions available to the player who moves after the history *In*, namely the incumbent, is $A(In) = Acquiesce, Fight$. Each player has his line of argument is called backward induction. Whenever a player has to move, he deduces, for each of the possible actions, the actions that other players will subsequently rationally take, and chooses the action that yields the most preferable terminal history.

A key concept of studying extensive games is strategy. A player strategy specifies the action he chooses for every history when taking the next move.

Definition (Strategy). A strategy of player i in an extensive game with perfect information is a function that assigns to each history h after which it is player i's turn to move (i.e. $P(h) = i$, where P is the player function) an action in $A(h)$ (the set of actions available after h).

Consider the game in the Figure 4.2.

- Player 1 moves only at the start of the game. The actions available to him are C and D. Thus, he has two strategies: one that assigns C to the empty history, and one that assigns D to the empty history.

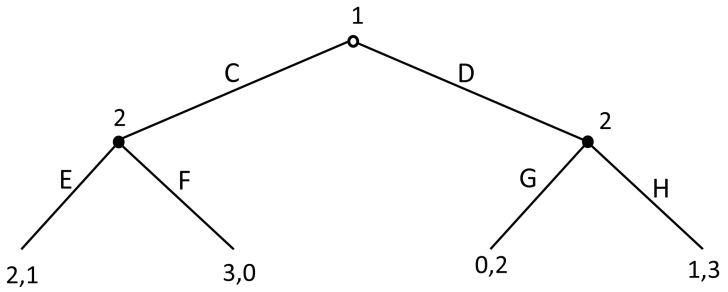


Figure 4.2: Extensive Form Strategies

- Player 2 moves after both the history C and the history D. After the history C the actions available to him are E and F, and after the history D the actions available to her are G and H. Thus a strategy of player 2 is a function that assigns either E or F to the history C, and either G or H to the history D. That is, player 2 has four strategies, which are shown in the Table 4.3.

	Action assigned to history C	Action assigned to history D
Strategy 1	E	G
Strategy 2	E	H
Strategy 3	F	G
Strategy 4	F	H

Table 4.3: Strategic form game

4.4.2 Nash equilibrium

One way to find a Nash equilibrium of an extensive game in which each player has finitely many strategies is to list each players strategy, find the outcome of each strategy profile, and analyze this information as for a strategic game. A Nash equilibrium: a strategy profile from which no player wishes to deviate, given the other players strategies. The definition is an adaptation of that of a Nash equilibrium in a strategic game. In other words, the notion of equilibrium in extensive games is refers to model the players' behavior that leads to a steady state. Hence, equilibrium is presented by patterns of behavior with the property that if every player knows every other player's behavior, he has no reason to change his own behavior.

Definition (Nash equilibrium of extensive game with perfect information). The strategy profile s^* in an extensive game with perfect information is a Nash equilibrium if, for every player i and every strategy s_i of player i, the terminal history $O(s^*)$ generated by

s^* is at least as good according to player i's preferences as the terminal history $O(s_i, s^*_{-i})$ generated by the strategy profile (s_i, s^*_{-i}) in which player i chooses s_i while every other player j chooses s^*_j. Equivalently, for each player i,

$u_i(O(s^*))u_i(O(s_i,\text{s}\hat{}*\dot{}\text{-i}))$ for every strategy s_i of player i,

where u_i is a payoff function that represents player i's preferences and O is the outcome function of the game.

Nash equilibrium yields a strategy profile from which no player wishes to deviate. The definition of Nash equilibrium is an adaptation of that of a Nash equilibrium in a strategic game.

- *Players.* The set of players in the extensive game.
- *Actions.* Each player's set of actions is his set of strategies in the extensive game.
- *Preferences.* Each player's payoff to each action profile is his expected payoff to the terminal history generated by that action profile in the extensive game. In conclusion, the set of Nash equilibria of any extensive game with perfect information is the set of Nash equilibria of its strategic form.

In the entry game in Figure 4.1, the challenger has two strategies, *In* and *Out*, and the incumbent has two strategies, *Acquiesce* and *Fight*. The strategic form of the game is shown in the Table 4.4. We see that the game has two Nash equilibria: *(In, Acquiesce)* and (Out, Fight). The first equilibrium is the pattern of behavior isolated by backward induction.

		Incumbent	
		Acquiesce	Fight
Challenger	In	2, 1	0, 0
	Out	1, 2	1,2

Table 4.4: Example: Entry game

In the second equilibrium the challenger always chooses *Out*. This strategy is optimal given the incumbent's strategy to fight in the event of entry. Further, the incumbent's strategy *Fight* is optimal given the challengers strategy: the challenger chooses *Out*, so whether the incumbent plans to choose *Acquiesce* or *Fight* makes no difference for the payoff. Thus, neither player can increase its payoff by choosing a different strategy, given the other players strategy.

The extensive game model a situation in which a challenger that always chooses *Out* never observes the incumbent's action, because the incumbent never moves. In a strategic game, the rationale for the Nash equilibrium condition that each player's strategy be optimal given the other players' strategies is that in a steady state, each player's experience playing the game leads his belief about the other players' actions to be correct. This rationale does not apply to the Nash equilibrium *(Out, Fight)* of the (extensive) entry game, because a challenger who always chooses *Out* never observes the incumbent's action after the history *In*. On rare occasions, non-equilibrium actions are taken (perhaps players make mistakes,

or deliberately experiment), and the perturbations allow each player eventually to observe every other player's action after every history. With this respect, it might happen that on those (rare) occasions when the challenger enters, the subsequent behavior of the incumbent to fight is not a steady state in the remainder of the game: if the challenger enters, the incumbent is better off acquiescing than fighting. That is, the Nash equilibrium *(Out, Fight)* does not correspond to a robust steady state of the extensive game.

Consequently, in order to address the problem of lack of robustness of Nash equilibrium for extensive games due to ignorance their sequential nature we introduce the subgame perfect equilibrium. Each player's strategy has to be optimal, given the other players' strategies, not only at the start of the game, but after every possible history.

Definition (Subgame). Let Γ be an extensive game with perfect information, with player function P. For any non-terminal history h of Γ, the subgame $\Gamma(h)$ following the history h is the following extensive game.

- *Players.* The players in Γ.
- *Terminal histories.* The set of all sequences h' of actions such that (h, h') is a terminal history of Γ.
- *Player function.* The player $P(h, h')$ is assigned to each proper subhistory h' of a terminal history.
- *Preferences.* Each player prefers h' to h'' if and only if she prefers (h, h') to (h, h'') in Γ.

Notice that the subgame following the initial history $\oslash$ is the entire game. Every other subgame is called a proper subgame. Because there is a subgame for every non-terminal history, the number of subgames is equal to the number of non-terminal histories.

In an equilibrium that corresponds to a perturbed steady state in which every history sometimes occurs, the players' behavior must correspond to a steady state in every subgame, not only in the whole game.

A *subgame perfect equilibrium* is a strategy profile s^* with the property that in no subgame can any player i do better by choosing a strategy different from s_i^*, given that every other player j adheres to s_j^*.

For example, the Nash equilibrium *(Out, Fight)* of the Entry game is not a subgame perfect equilibrium because in the subgame following the history *In*, the strategy *Fight* is not optimal for the incumbent: in this subgame, the incumbent is better off choosing *Acquiesce* than it is choosing *Fight*. The Nash equilibrium *(In, Acquiesce)* is a subgame perfect equilibrium: each players strategy is optimal, given the other players strategy, both in the whole game, and in the subgame following the history *In*.

Definition (Subgame perfect equilibrium) The strategy profile s^* in an extensive game with perfect information is a subgame perfect equilibrium if, for every player i, every history h after which it is player i's turn to move (i.e. $P(h) = i$), and every strategy s_i of player i, the terminal history $O_h(s^*)$ generated by s^* after the history h is at least as good according to player i's preferences as the terminal history $O_h(s_i, s_{-i}^*)$ generated by the strategy profile (s_i, s_{-i}^*) in which player i chooses s_i while every other player j chooses

s_j^*. Equivalently, for every player i and every history h after which it is player i's turn to move,

$u_i(O_h(s^*)) \geq u_i(O_h(s_i, s*_i))$ for every strategy s_i of player i,

where u_i is a payoff function that represents player i's preferences and $O_h(s)$ is the terminal history consisting of h followed by the sequence of actions generated by s after h.

In a subgame perfect equilibrium every player's strategy is optimal, in particular, after the initial history (put $h = \oslash$ in the definition, and $O_\oslash(s) = O(s)$). A subgame perfect equilibrium generates a Nash equilibrium in every subgame: if s^* is a subgame perfect equilibrium then, for any history h and player i, the strategy induced by s_i^* in the subgame following h is optimal given the strategies induced by s_{-i}^* in the subgame. Further, any strategy profile that generates a Nash equilibrium in every subgame is a subgame perfect equilibrium, so that we can give the following alternative definition. A subgame perfect equilibrium is a strategy profile that induces a Nash equilibrium in every subgame.

We consider again the entry game in the previous example, which has two Nash equilibria, *(In, Acquiesce)* and *(Out, Fight)*. The fact that the Nash equilibrium *(Out, Fight)* is not a subgame perfect equilibrium follows from the formal definition as follows. For $s^* = (Out, Fight)$, $i = Incumbent$, $s_i = Acquiesce$, and $h = In$, we have $O_h(s*) = (In, Fight)$ and $O_h(s_i, s_{-i}^*) = (In, Acquiesce)$, so that the inequality in the definition is violated: $u_i(O_h(s^*)) = 0$ and $u_i(O_h(s_i, s_{-i}^*)) = 1$.

The Nash equilibrium *(In, Acquiesce)* is a *subgame perfect equilibrium because* (a) it is a Nash equilibrium, so that at the start of the game the challenger's strategy *In* is optimal, given the incumbents strategy *Acquiesce*, and (b) after the history *In*, the incumbent's strategy *Acquiesce* in the subgame is optimal. In the language of the formal definition, let $s^* = (In, Acquiesce)$. The challenger moves after one history, namely $h = \oslash$. We have $O_h(s^*) = (In, Acquiesce)$ and hence for $i = challenger$ we have $u_i(O_h(s^*)) = 2$, whereas for the only other strategy of the challenger, $s_i = Out$, we have $u_i(O_h(s_i, s_{-i}^*)) = 1$.

The incumbent moves after one history, namely $h = In$. We have $O_h(s^*) = (In, Acquiesce)$ and hence for $i = incumbent$ we have $u_i(O_h(s^*)) = 1$, whereas for the only other strategy of the incumbent, $s_i = Fight$, we have $u_i(O_h(s_i, s_{-i}^*)) = 0$. Every subgame perfect equilibrium is a Nash equilibrium, so we conclude that the game has a unique subgame perfect equilibrium, *(In, Acquiesce)*.

A Nash equilibrium of a strategic game corresponds to a steady state in an idealized setting in which the participants in each play of the game are drawn randomly. The idea is that each player's long experience playing the game leads him to correct beliefs about the other players' actions; given these beliefs the equilibrium action is optimal. A subgame perfect equilibrium of an extensive game corresponds to a slightly perturbed steady state, in which all players, on rare occasions, take non-equilibrium actions, so that after long experience each player forms correct beliefs about the other players' entire strategies, and thus knows how the other players will behave in every subgame. Given these beliefs, no player wishes to deviate from his strategy either at the start of the game or after any history. This interpretation of a subgame perfect equilibrium, like the interpretation of a Nash equilibrium as a steady state, does not require a player to know the other players' preferences, or to think about the other players' rationality.

5 Deliberation Forum

Deliberation, i.e., discussing and ranking different proposals and making decisions, is an important issue for many communities, be they political, be they boards of experts for a scientific issue. Online deliberation however has issues, such as unorganized content, off-topic or repetition postings, or aggressive and conflicting behavior of participants. To address these issues, based on a relatively simple argumentation model and on feedback of different type, 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.

The question we want to investigate is whether a simple, intuitive argumentation model, but together with ratings by participants, possibly of different type, allows to identify useful points, arguments and convincing proposals. In a nutshell, the question 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. Our perspective is that, in such contexts, vast range of interests and priorities of individuals affect the outcome much more than argument quality. Furthermore we hypothesize that making the community deliberate on the issues in question should increase its satisfaction with subsequent decisions if the scheme that identifies and selects solutions to the issues discussed takes this deliberation into account. In the end, the question investigated here is how online deliberation can be organized so that satisfying decisions can be derived from it.

In this chapter we introduce a deliberation forum model that supports bringing structure in the discussion and comprehensive evaluation. Our approach consists of three steps: (1) assigning weights to participants based on the set of formal criteria such as degree of engagement in the discussion to stimulate desirable behavior , e.g., originality of arguments, focus on the topic in question etc.; (2) assigning scores to comments, considering the weights of authors and raters and the agreements/disagreements of the community with the expressed argumentation; (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. Next, given such a scoring scheme, it is important to examine to which extent individuals have understood and accepted the approach, to identify characteristics of good discussants and of strong arguments and proposals, and to study the robustness of the approach with regard to minor changes.

To this end, we have built a respective online platform and 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 [124, 125]. 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. Our takeaway is that the approach proposed here is promising to improve deliberation in many settings.

Our contributions presented in this chapter are as follows: First, we list design decisions and motivate each one of them giving their implications on our forum model. Next, 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. We incentivize such favorable behavior by giving participants different degrees of influence on the evaluation of the argumentation, contingent on their weights. As a next step, we introduce the scoring scheme that for each argument assigns 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 scoring scheme. Our scheme assigns each proposal a score that reflects the share of pro and contra arguments and their scores. Then we present the detailed formalization of the model with the technical details and related formulae. Finally, we describe the evaluation settings, the experiment we have conducted with more than 100 participants. We close the chapter with the results and the conclusion.

5.1 Design decisions

To shed more light on our approach, we list our main design decisions in this section.

The look-and-feel of our deliberation forum is the one of a conventional forum wherever possible. Design and interaction features of our forum are mainly the ones of a classical forum. An alternative would have been an entirely new design. However, user acceptance is critical in our context, and our alternative is likely to be better in this respect. Further, we can leverage existing technology and the host of comfort features provided by current implementations.

Comments are typed, and the typing mimics common argumentation structures. We have introduced comment types such as pro and contra arguments. In our forum, each proposal corresponds to a separate thread containing the respective arguments, so that their discussions are clearly separated from each other.

The forum model should be simple and intuitive. We value simplicity of the model more than exactness and comprehensiveness. However, instead of having a model that is highly comprehensive but overly complicated for non-experts, and familiarization with it requires time and effort, we have limited our model to elementary comment and rating types.

Community members have different weights, according to formal criteria. In order to incentivize community members following our guidelines when deliberating, we have decided to assign them weights, and a higher weight gives an individual more influence on the decisions which will be taken. Next, we have decided to let the weight of individuals depend solely on formal criteria (such as number of arguments provided, share of arguments not flagged as repetitions by the community) subsequently referred to indicators. While each argument by an individual meets a certain degree of agreement in the community, the

weight of an individual does not depend on the degree of agreement of his arguments. The rationale is not to discriminate against minority opinions.

The weight of a participant is the minimum of all his indicator values. An indicator value reveals the extent of a participant obeying the respective formal criterion. We deem it important that participants observe all of our criteria, and we want to incentivize such behavior. To illustrate, we do not want to give a high weight to an individual who has issued many arguments if the community labels relatively many of them as off-topic. Hence, the weight is the minimum function of all indicator values.

Individuals can give feedback on contributions by others, feedback is typed, and it is used according to its type. Participants may issue feedback on various contributions by others, which are then used in different ways. For instance, participants can state that they agree or disagree with an argument issued by someone else or can mark it as off-topic or as a repetition of a previous argument. An alternative would be to combine the various ratings/feedback items of different type into one argument score. However, we have found this too undifferentiated. For instance, agreement/disagreement ratings are used to quantify the degree of acceptance of the community, while off-topic/repetition feedback is used to reject comments.

Participants can see their own weight. The alternative would be to not show this information so that participants are not influenced by it. Our decision has been to display current indicator values to give the participants an idea how their behavior so far has affected their weights. The rationale has been that this might stimulate the behavior desired.

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.

We evaluate our approach experimentally. An alternative to experiments would have been a formal analysis of the approach or an evaluation with a simulation. 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.

5.2 Features

As stated before, a design objective of ours has been that the platform envisioned has the interface of a conventional discussion board as much as possible and incorporates its functionality. In addition to the usual parts of online discussion models (posts and references to previous posts), our approach has several new features: (1) **comment types**, i.e., an author is supposed to categorize his content according to our argumentation model, e.g., pro argument or contra argument comment; (2) **multi-facet ratings**, e.g., participants can give their feedback on whether they agree or disagree with the content of a comment and rate its

writing style, tone, or type. Quite naturally, these features require some modifications of the look-and-feel of a conventional discussion board. Further constituents of our approach are as follows: (3) a **weighting scheme** for participants reflects their adherence to our formal criteria for constructive deliberation and rewards them accordingly with different degrees of influence in the discussion; (4) a **scoring scheme** for the evaluation of comments; (5) a **scoring scheme for proposals** based on the argumentation presented. In the subsequent section, we describe these features in detail and motivate them.

5.2.1 Discussion structure

To address the issue of unorganized content in forum discussions, we have proposed a structure based on topics (discussion issues), solutions or proposals how these issues can be solved and in the end, comments as arguments in favor or against a particular solution proposal. 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, solution 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, related, thus, to a solution. Comments are typed, e.g., pro argument or contra argument. A comment can also 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.

In the figures 5.1 and Figure 5.2 are shown screenshots of our deliberation forum and its structure. First figure refers to discussion topics, whereas the second one is the overview of existing proposals for one of the topics.

5.2.2 Comment types

To facilitate categorization of a post based on its argumentation and to support evaluation of the arguments, authors can classify comments in different types, namely proposal, proposal extension, pro argument, contra argument or other, according to the argumentation model we are proposing. Thus, we have slightly adjusted the discussion structure of conventional discussion boards to encompass these various categories, as follows:

A proposal represents an idea or suggestion how a discussion issue can be solved. In our study with a community of computer-science students, there have been several discussion issues such as: "How to spend a EUR 500,– budget on behalf of the students?" or "Which student should an iPad be given to?". Each discussion issue forms a separate forum thread. The number of proposals for each discussion issue, which represent separate proposal threads

Figure 5.1: Deliberation forum - discussion topics

within a forum thread, is arbitrary. Examples of proposals regarding one of the issues just mentioned have been following: The amount of EUR 500,– can be spent to support a project for the live streaming of lectures; or Improve WLAN at important spots on the university campus.

A new proposal extension is a comment referring to proposal suggesting some improvements/extensions. I.e., one of the proposals for the issue: Which student should an iPad be given to?The proposal has to relate the chance of winning the iPad with exam points, and a proposed extension has been to organize a lottery and assign lots proportional to the number of exam points earned.

A pro comment is a comment which contains argumentation in favor of a certain proposal.

A contra comment is a comment containing arguments/reasons against a certain pro-

Figure 5.2: Deliberation forum - proposals

posal.

A comment of type *Other* is a comment which does not match the categorization just presented, or its author does not want to assign it to one of these categories.

The comment types presented and the underlying argumentation model have affected the discussion structure as follows. Discussion issues form separate **forum threads**. Within each forum thread there are different **proposal threads**. Authors post comments in proposal threads directly referring to the discussed proposal. This is the case even when they are posted as follow-ups of other comments.

In the Figure 5.3 are shown these different types of comments available. Notice that the labels are in German: "Erweiterung des Vorshlags" (*Proposal Extension*), "Neues Pro

Figure 5.3: Deliberation forum - comment types

Argument..." (a new *Pro comment*), "Neues Gegenargument..." (a new *Contra comment*) and Sonstiges (*Other*).

5.2.3 Multi-facet ratings

Our setting allows for feedback addressing different characteristics of a comment (of any type):

Content. The object of a content rating is the content of a comment, and the following values are possible: *Agreement*, *Disagreement*, *Repetition*, *Off-topic* and *Other*. A rater can express his agreement or disagreement with an argument expressed or address structural characteristics of a comment, e.g., marking it as repetition or off-topic.

Writing style. A rating for writing style reflects how a comment is written on the grading scale from 1 to 5, e.g., Rate 5 stands for clear and concise writing, Rate 1 for unclear, fragmentary input.

Tone. Tone ratings address the tonality of comments on the grading scale between 1 and 5. Rate 5 corresponds to comments which are balanced and polite, whereas provocative and confrontational comments are rated with 1.

Comment type. The object of a comment type rating is the comment type. In order to verify comment types, raters are invited to classify the argumentation of comments themselves. The possible values are *Proposal extension*, *Pro comment*, *Contra comment*, and *Other*.

Figure 5.4 gives overview of these different feedback options. Notice that the labels are in German: "Ihre Bewertung..." (*content rating*), "Stil der Argumentation..." (*writing style*), "Hoefflichkeit" (*tone*) and "Ihre Typisierung..." (*comment type*). "Zustimmung", "Ablehnung", "Wiederholung", "Off-topic" and "Sonstiges" stand for *Agreement*, *Disagreement*, *Repetition*, *Off-topic* and *Other*.

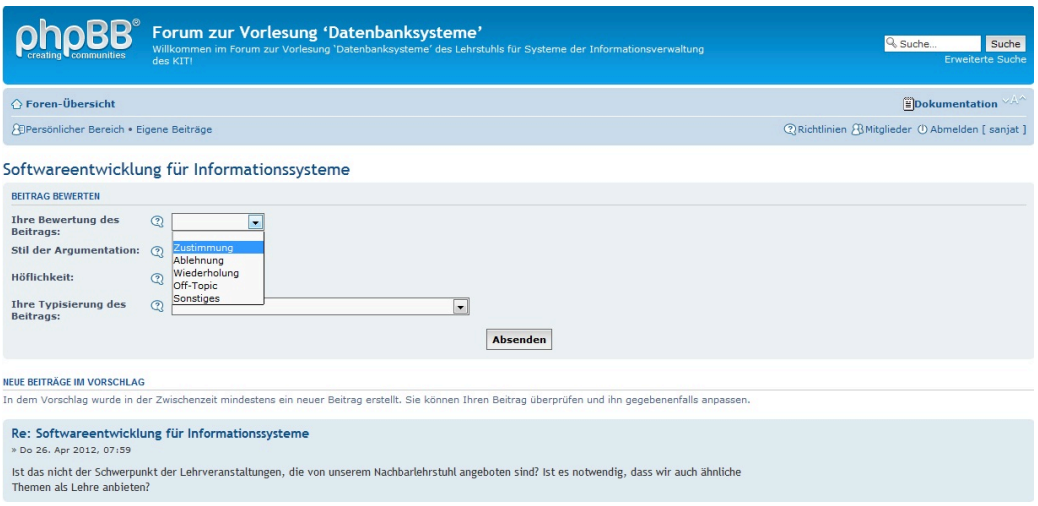


Figure 5.4: Deliberation forum - different types of feedback

5.2.4 Weighting scheme

In the forum model we have envisioned a core objective is to facilitate constructive deliberation without repetitions, off-topic comments, offensive and harsh behavior. To this end, we propose formalizations of unwanted behavior. So-called indicators quantify the degree of adherence of a participant to each criterion. As an incentive to refrain from such behavior, we confine the influence of participants with such behavior in the forum. Our list of criteria is as follows:

Originality. A participant performs well regarding this criterion if he has posted no or very few repetitions of already existing comments. This criterion should decrease repetitions in the discussion.

Focus. The value of this indicator will be high if a participant has received no or a very few off-topic ratings for his comments. Our motivation is to lower the number of off-topic comments and make the discussion more efficient.

Style. If raters rate the writing style of the comments authored by an individual high, i.e., rates 4 and 5 on the grading scale with 5 as maximum, this will affect his performance regarding this criterion. The rationale is to increase the clarity of comments.

Tone. The better the tone of an author is rated, the higher will be his value for the Tone criterion, similarly to the style criterion. The aim with this criterion is to keep the discussion friendly and balanced.

Engagement. The value of this criterion will be larger, the larger the number of posted comments and ratings by the author in question is. The rationale is to reward above-average active discussants.

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. The higher the engagement of a participant in different discussions, the higher is the value of the *Breadth* criterion. Our perspective is that participants with broad interests should be rewarded. Additionally, this criterion should make it more difficult for groups of individuals with specific, narrow interests to collude.

Honesty. The value of the *Honesty* criterion is larger, the larger the score of participant ratings, the so-called hfmscore, as assigned by the so-called peer prediction method [126]. 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*. The aim of this and similar approaches [127] is to maximize the reward for honest answers in the absence of an objective truth criterion. Since ratings are an important part of our approach, a mechanism to assess them is needed.

Our objective with this current study is not to arrive at a list of criteria that is final and covers all aspects of desirable behavior in online deliberation. (Instead, we have aimed at coming up with one concrete proposal and then evaluate it subsequently.) Nevertheless, we hypothesize that we have identified the most important points for constructive and efficient discussions, taking into account problems that previous projects have faced.

5.2.5 Formulae and notations

Here we give a more rigorous introduction to our scoring scheme. First, to illustrate, we will elaborate on the formal criteria used to assign weights to participants, along with our formula for calculating this weight. Next, we introduce our formulae for calculating comments' and proposals' scores.

Weighting scheme. Our weighting scheme relies on eight different indicators, as described informally in the previous subsection.

As stated already, our approach features ratings of different type. A rating consists of rates for: content {*Agreement*, *Disagreement*, *Off-topic*, *Repetition*, *Other*}, writing style and tone, both presented by a grading scale between (1−poor) and (5−good), comment type *Pro comment*, *Contra comment*, *Proposal extension*, *Other*.

R is the set of all ratings. $R(k)$ is the set of ratings posted for Comment k, and $R^{create}(j)$ is the set of ratings posted by Participant j. The set of all ratings posted for comments of Author j is $R^{subject}(j)$, analogously, $R^{subject}_{off-topic}(j)$, $R^{subject}_{repetition}(j)$ are the sets of off-topic, repetition ratings, respectively.

Focus. The indicator focus is defined as the ratio of the number of all off-topic ratings received for comments posted by Participant j over the same number of all ratings received.

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

Originality. The originality indicator is calculated mainly based on the share of off-topic ratings referring to comments issued by Participant j compared to all ratings referring to these comments.

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

Accordingly, $R^{subject}_{style-}(j)$, $R^{subject}_{tone-}(j)$ are the sets of low ratings (Rate 1 and 2 on the grading scale 1 to 5) for style and tone for comments authored by Participant j respectively.

Style. The style indicator is calculated as follows:

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

Tone. The formula for the tone indicator of Participant j is as follows:

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

$K^{create}(j,t)$ is the set of all useful comments posted by Participant j in discussion thread t. Similarly, $K^{create}(j)$ is the set of useful comments by j in all threads. A useful comment is one that has less than 50% of off-topic or repetition ratings. P is the set of all participants.

Engagement. This indicator is calculated based on the number of ratings and comments issued by Participant j compared to average numbers over all participants.

$$engage(j) = \frac{\left|K^{create}(j)\right|}{avg_{i \in P}(|K^{create}(i)|)} + \alpha_{engage} \cdot \frac{\left|R^{create}(j)\right|}{avg_{i \in P}(|R^{create}(i)|)}$$

Individuality. To compute the indicator for individuality, we rely on the following auxiliary measures: similarity of posting and of rating behavior of participants. Both of these measures rely on opinion. For instance, participants who agree on the same comments or who post pro comments for the same proposal have a similar opinion. $K^{pro}_{simil}(i,j,t)$ is the maximum of the number of pro comments authored by Participant i in Thread t and of the number of such comments authored by Participant j. $K^{contra}_{simil}(i,j,t)$ is defined analogously for contra comments. $K_{simil}(i,j)$ is the sum of those numbers over all threads. $K_{disimil}(i,j)$ is defined analogously. $R_{simil}(i,j)$ is the set of tuples of ratings ($r1$, $r2$) posted by Participant i and j for a comment that are both either agreement or disagreement. $R_{disimil}(i,j)$ in turn is the set of all tuples of ratings expressing different opinions for a comment issued by Participants i and j. Consensus and contention of Participants i and j now is defined as follows:

$$cons(i,j) = \frac{|R_{simil}(i,j)| + |K_{simil}(i,j)|}{|R_{simil}(i,j)| + |K_{simil}(i,j)| + |R_{disimil}(i,j)| + |K_{disimil}(i,j)|}$$

$$noncons(i,j) = \frac{|R_{disimil}(i,j)| + |K_{disimil}(i,j)|}{|R_{simil}(i,j)| + |K_{simil}(i,j)| + |R_{disimil}(i,j)| + |K_{disimil}(i,j)|} \quad .$$

Next, based on consensus and contention of Participant j with other participants, we find, the set of participants that sometimes match and sometimes do not match the opinion of j.

$$P^{partlyDiff}(j) = \{i \in P : cons(i,j) > 0.3 \wedge noncons(i,j) > 0.3\}$$

The value 0.3 was set empirically, based on the results of a preliminary experiment. We aimed at nudging discussants to deliberation considering pros and contras. Although there is a certain arbitrariness in choosing the threshold, the data analysis described subsequently has shown that altering this value does not have a significant effect on the experiment outcome. Finally, the individuality indicator is the ratio of these individuals over the whole set of participants.

$$indiv(j) = \frac{\left|P^{partlyDiff}(j)\right|}{|P|}$$

T is the set of all forum threads, and $T^{create}(j)$ is the set of all forum threads Participant j has actively participated in. Active participation is given if the number of useful posts is at least half of the average number of useful posts by all participants in that forum, as formalized below. We have chosen a value lower than the average number of comments for this threshold, in order to be less discriminative on the number of comments and still honor participants who actively take part in the discussion.

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

Breadth. The value for breadth is the ratio of the number of forum threads where the participant has actively taken part in and the total number of forum threads:

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

Honesty. The honesty indicator is calculated using scores assigned by the peer prediction method [126]. Based on the probability distribution of the given rating and the scoring function, a score for the rating is assigned accordingly. $hfmscore(j)$ is the average of all rating values issued by Participant j.

The first four indicators (*originality*, *focus*, *style*, *tone*) are not normalized. These indicators refer to a minority behavior such as posting repetitions or off-topic comments, and the rationale behind not normalizing is to demarcate minority behavior from regular behavior. On the other hand, *engagement*, *individuality*, *breadth* and *hfmscore* are normalized based

on frequency. For instance, if 20% of the community has performed better than Participant j regarding the breadth criterion, j's normalized value of the breadth indicator is 0.8. The advantage of this normalization is that the distribution is uniform in the range [0, 1], and values for different criteria now are comparable.

The weight of a participant is the minimum of the different indicator values. We could have used another function here as well, but we have decided to use minimum function. This is to reward participants with higher weights when they obey all criteria.

$$weight(j) = \min(focus(j), orig(j), style(j), tone(j), \\ breadth^{norm}(j), engage^{norm}(j), indiv^{norm}(j), hfmscore^{norm}(j))$$

Comment scoring scheme. To assess comments and their argumentation we propose a respective scoring scheme. It takes the following criteria into account: author weight, rater weights and agreement status in the community, e.g., the ratio of agreement ratings received and all ratings. In what follows, k is a comment, $weight(author(k))$ is the weight of its author, and $weight(issuer(r))$ is the weight of the individual who has generated rating r. $R_{ref}(k)$ 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 Comment k. Accordingly $R^{opinion}_{ref}(k)$ is the set of all ratings of type *Agreement* and *Disagreement* for Comment k. $K(F)$ is the set of all comments in Forum F.

$$score(k) = \left(\frac{weight(author(k)) + \sum_{r \in R^{+}_{ref}(k)} weight(issuer(r))}{weight(author(k)) + \sum_{r \in R^{opinion}_{ref}(k)} weight(issuer(r))} - 0.5 \right) \cdot w_1(k)$$

where

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

Comment scores are normalized using Weight $w1$. It is the ratio of the sum of weights of the author and the raters of Comment k and the maximum sum of weights of author and raters for any comment in the Forum F. The rationale is to normalize, i.e., to make comment score comparable on the forum level.

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

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. To make proposal scores mutually comparable in a forum, they are normalized as well. We accomplish this by using the maximum difference between pro and contra arguments of a proposal in that particular forum. 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.

5.3 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.

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 indicators' calculation. Having engaged students of a database course, the motivation is justified by their technical background.

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. Participants' weights are important part of our decision-making scheme and they have a strong impact on the scores.

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. The hypothesis reflects a proposition that a properly designed scheme can increase the tolerance of the community to the decisions.

H5: The higher the evaluation of the decision-making scheme, the higher is the perceived quality of the decisions. This claim is similar to the previous hypothesis.The distinction is that the more useful a scheme is perceived, the higher will be the perceived quality of the decisions.

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.

H7: The degree of adherence to our criteria is positively correlated with the fairness perceived. The rationale behind this hypothesis is to gather further evidence whether our design works in the way we have expected. If participants consider proposed formal criteria fair, it will be easier for them to behave in accordance with them.

H8: Displaying weights of participants affects their behavior. In other words, the indicator values displayed to all participants will influence their behavior in a way that is desirable.

The above presented list of hypotheses was a starting point for designing a questionnaire that could reveal answers to these research questions based on participants' reports.

5.4 Experimental setup

To evaluate our approach we have conducted an experiment. Alternatives would have been to perform formal analysis or simulations. A downside of these alternative methods is that they require certain simplifications and assumptions such as the distribution of posts or ratings etc., which are unknown at this point. On the other hand, experimental studies are not easily repeatable at all, for various reasons: Recruiting a community is difficult and time-consuming, and if the experiment was repeated with another parameter setting, new issues to be discussed would have to be identified. Further, solutions/decisions drawn from the discussion need to be effectuated in order to make the experiment as realistic as possible, and this tends to be costly.

The community, we have conducted our study with is second-year computer-science students of our university. We have run an experiment with around 200 participants. The experiment was running for four weeks. In order to incentivize participation in principle, we have announced a small bonus for the final exam of the database course, which is mandatory in the fourth semester (5% of the exam points that could be earned in total for five comments, no repetitions or off-topic comments and 20 ratings). 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. Beyond that, we have announced that decisions are binding (for us, the organizers of the experiment), i.e., we will implement the winner proposals. For instance, we have promised that we will indeed offer the course with the highest degree of agreement in the subsequent academic year, and that we will try hard to find a lecturer in case we are not able to teach it ourselves (analogously with the iPad or the EUR 500). Regarding the issues where the winner proposal is brought to the attention of the dean, we also deem this to be a real incentive, since it should be of interest to the management of the department to get to know the preferences of an entire age group. The

discussion itself has lasted four weeks. After that, we have indeed, say, spent EUR 500,– (to support a project for the video streaming of lectures) and offered a course on NoSQL databases.

The discussion topics have been as follows:

What should be the topic of the last session of the current database course? We have proposed three different lecture topics, which participants then discussed.

How should a budget of EUR 500,– be spent on behalf of the students? We have required that the money must be spent as a whole and in a way that does not violate German regulations for spending public money.

We have offered a new iPad to be given away according to a criterion proposed by the community. A proposed criterion should be objectively measurable, so a proposal such as 'John Doe' is not acceptable in this respect.

Assuming that the computer-science department had funding for a new chair, what should be its research direction?

What should be the topic of a new course in the area of database/information systems in the next academic year?

Considering the existing selection criteria for the KIT master program in computer science, which one should be given higher priority?

What is the most urgent reform of the KIT bachelor program in computer science?

On the technical level, we have developed the discussion portal as an extension of the well-known open-source forum software phpbb [1]. phpBB is listed in relevant blogs and forums as one of the top ten open-source forum projects. It is originally written in php and supports various database systems such as MySQL, which we have used. Extensions we have introduced are new interface features such as specifying the type of a comment; multi-facet rating options, e.g., content/writing style ratings. See our forum [2]. Other important new features are the implementation of the weighting scheme to profile the influence of participants in the discussion and the scoring scheme described here (see 5.2.5). Additionally, we have adapted the interface in order to anonymize the data, i.e., we did not want to display information such as the names of comment authors.

We have decided to evaluate our forum model by means of a questionnaire (see Appendix .1). 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 how to spend a budget of EUR 500, since it fosters socializing within that community even though the organizers of this experiment might not like it. Further, our research question has been how to arrive at decisions satisfying to most of the community members and not necessarily at good decisions. 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

[1] https://www.phpbb.com/
[2] http://shakuras.ipd.uni-karlsruhe.de/dbsforum/

relate questionnaire answers to user behavior in our system. The detailed analysis of the user data collected during a 4-week-experiment is a subject of subsequent section.

5.5 Questionnaire 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. Primarily, we have been interested in whether participants have understood the discussion structure, namely different types of comments and feedback, and adopted them. Furthermore, it was very important to get insights into their opinion regarding fairness of the weighting and scoring schemes. The outcomes stem from verifying the following hypotheses:

H1: Participants have deemed our weighting scheme fair. In absolute numbers, only 19 participants out of 116 participants have rated the fairness of the model as moderate, i.e., Rates 1 (not fair at all) and 2 (not fair). 31 participants were neutral. The highest positive correlation between the perception of the fairness of the weighting scheme and the fairness of the individual 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. According to the questionnaire data, there is a significant correlation ($r = 0.5433$, $p < 0.001$). Out of 116 participants, only 11 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 the connection between weights and the way proposals and arguments are scored.

H4: The higher the perceived usefulness of decision-making scheme, the more satisfied is the community with the winner proposals. Although we have not confirmed a significant correlation, participant ratings of the usefulness of the decision-making scheme have been 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 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 statistical point of view.

H6: The perceived quality of the decisions is positively correlated with the participants' feeling that their opinion is respected. There is a significant correlation

($r = 0.2327$, $p < 0.02$). Participants reported that the quality of the final decision is closely related to the perceived respect of the opinion in the forum.

H7: The degree of adherence to our criteria is positively correlated with the fairness perceived. We have not discovered any significant correlation. Note that this does not mean that the relationship that forms the basis of the hypothesis does not exist; it is just that we have not been able to validate it in our setup and with our questionnaire. In absolute numbers, 31 participants out of 116 have stated that the criteria and the weights have influenced their behavior in the forum. Refer to the following subsection Discussion for a respective discussion.

H8: Displaying weights of participants affects their behavior. Again, we cannot confirm this hypothesis here. Only 13 participants out of 116 have stated that their weights or the ones of their peers have affected them. We stress that we have communicated our criteria in detail; still, according to most participants, this has not influenced their behavior. Again, see the following subsection Discussion for more details.

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

To summarize, an important insight of the questionnaire results is that participants have been positive about many aspects of our approach. On the other hand, we could not confirm all of our hypotheses in this particular setup. More specifically, 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 their 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.

5.6 Discussion

Although the questionnaire results have been helpful to answer some of our questions, there are some results that leave room for different interpretations.

5.6.1 Questionnaire results

Looking at the free-text answers in the questionnaire, 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. According to our web-access statistics, the majority of participants has not fully read the documentation of the weighting and the decision-making scheme. E.g., participants have evaluated criterion *Honesty* high, although most of them probably have not understood the peer-prediction

method, and a few individuals have complained about their scores. Another example is that the decision-making scheme, though rated highly, has not yielded more acceptable decisions from the community point of view. The participants have acknowledged that respect of opinions of others is higher, and that decisions are of higher quality. Still, the tolerance towards decisions taken did not appear to be higher. Furthermore, participants rarely have given full-text answers containing constructive comments on how the model could be improved, or explaining why they have not been satisfied.

5.6.2 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. However, when trying to convince another community to adopt our approach in order to come to decisions, there has been some resentment that our scheme was 'not democratic' because of that reason. However, our approach does not violate the principle of equality according to the German constitution since it treats all participants equally; we have consulted with legal experts on this issue. Further, our perspective is that the criteria are clear and well-documented. In addition, 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. Further, participants with a low weight can still influence the decisions, by coming up with arguments that are well received by most community members.

5.6.3 Forum model

The motivation behind our work has been to foster deliberation and to give way to decisions widely accepted by the community. We have conducted our evaluation with the audience of a university course. This has some differences to other communities: First, an age group 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. For instance, think of public or political discussions which gather different groups of individuals regarding motivation, interests, educational and social background. Second, in our context, while bonus points have been an important incentive, they have certainly not been the only stimulus for participation. Students have shown interest for 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. Finally, our rules for earning the bonus points have affected the behavior of participants. They should 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, the satisfaction rate has been high, as described earlier. This gives way to the expectation that our approach will also work in settings without any external incentives such as bonus points.

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.

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

As stated before, the evaluation of proposal extensions is an open issue which is very difficult to solve, 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. Addressing these questions exceeds the scope of this current study and is future work. 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.

5.7 Data Analysis

Beyond the questionnaire results, analysis of the data collected during the experiment is to give us a broader perspective in terms of performance of our approach and its robustness even in the absence of an objective truth criterion. Here we present results from the conducted analysis of the experimental data [125]. The structure of what follows is in line with the main points of our approach: weighting scheme, then scoring schemes for comments and proposals. To begin with, however, to gain further insight in the discussion board, led discussions and board community, the posting behavior plays a significant role. These fundamentals can acknowledge the performance of our approach in a setting very close to the real world. Additionally, the analysis of posting behavior can provide indications of collusion or misuse attempts.

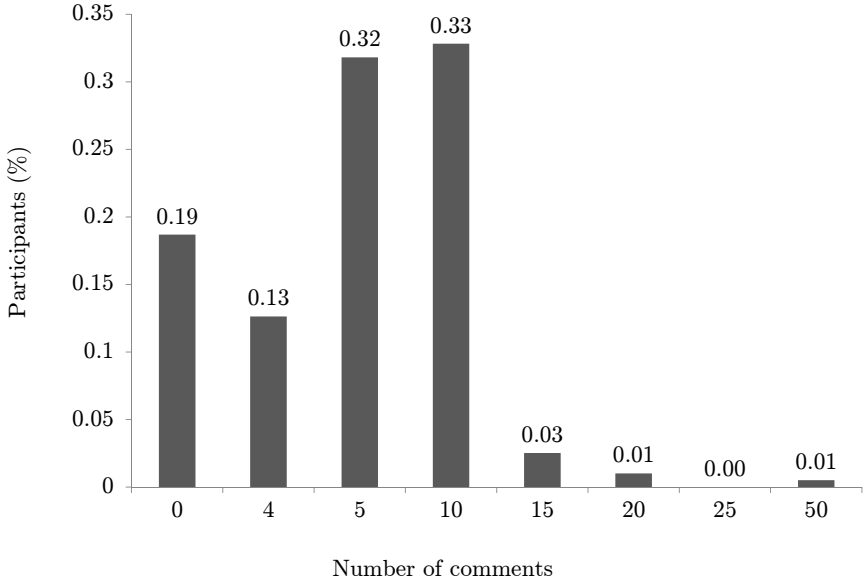


Figure 5.5: Distribution of comments

5.7.1 Posting behavior

Number of Contributions. In the four-week experiment, 954 posts and 3849 ratings were generated. 198 participants were registered, and 169 (84%) have posted at least one comment or one rating. When we observe registration of participants, we can conclude that the distribution of registrations has peaks at the beginning of the discussion period and at one point of time when we announced that the discussion period was extended. Half of the total number of participants has registered in first four days of the experiment. Still, if we compare early registered users to later ones, the distribution of posts and ratings is quite uniform.

Figure 5.5 is a histogram of the share of registered participants with a certain number of comments posted. 70% of these participants posted five or more comments, whereas 38% of participants posted more than five comments (which has been the limit for receiving the small exam bonus in full). Similarly, as presented in Figure 5.6, 67% of the participants posted 20 ratings or more, and 55% of the participants posted more ratings than required for receiving the bonus. So, we can confirm that there has been a certain intrinsic motivation for participation.

Posting Behavior. According to the experimental data, we can confirm a significant correlation between the number of posts and the number of ratings per participant ($r = 0.5748$, $p < 0.0005$). Additionally, we have discovered a significant correlation between the number of follow-up comments, i.e., comments referring to other comments, and the number

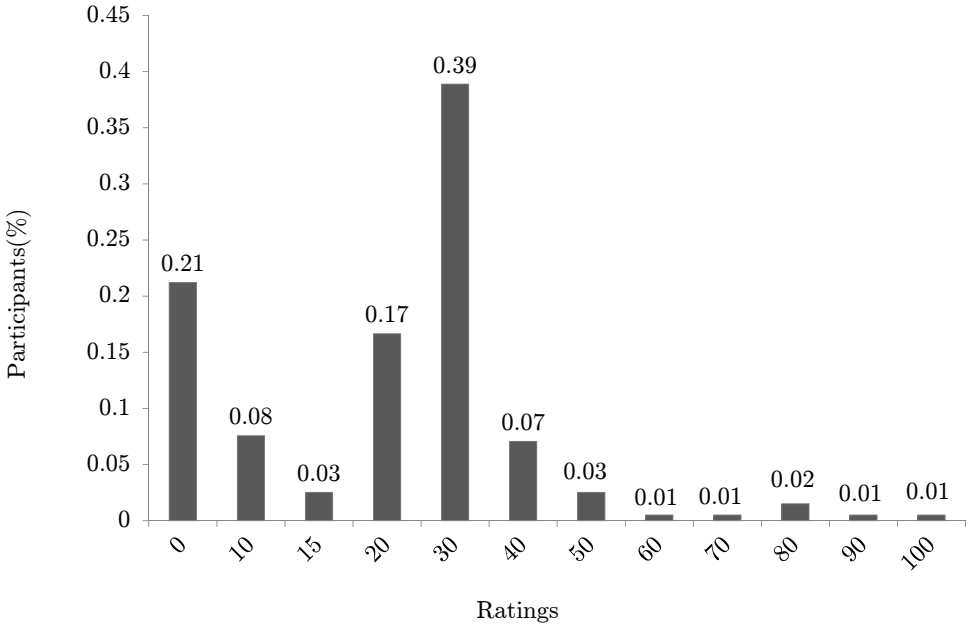


Figure 5.6: Distribution of ratings

of ratings ($r = 0.3821$, $p < 0.001$). In other words, participants who posted more comments have also posted more ratings and more follow-up comments.

In the experiment, 161 participants (81% of registered participants) posted at least one comment. Out of a total number of 954 comments there are 88 *Proposals* (9.22% of the comments), 162 *Proposal extensions* (16.98%), 84 *Other* comments (8.8%), 241 *Pro comments* (25.26%), 379 *Contra comments* (39.73%). The type *Other* is rarely used, and this speaks in favor of the acceptance of our argumentation model. Furthermore, the participants who had more posts also had more posts of type *Other*, and the type is used by 54 different participants. The correlation between the number of posts per participant and the number of *Other* posts is significant, $r = 0.4668$, $p < 0.001$. The number of *Other* comments is fairly small compared to the total number of comments. On the other side, more than a quarter of all registered participants posted them, so it is not confined to a small group of participants. A possible explanation is that participants fail to categorize these comments in rare cases. Thus, according to our interpretation, they are exceptions, rather than indications of misuse attempts or of participants having misunderstood some underlying notions.

We have observed a significant correlation between the number of pro and contra comments by the date of post ($r = 0.886057$, $p < 0.001$). This correlation indicates that participants were involved in the discussion in a differentiated manner, responding to arguments with pro and contra arguments.

Rating Behavior. We now look at the number of ratings per post. In total there are

3849 ratings posted by 156 participants. Out of these, 2364 (61.42%) are agree ratings, 1058 (27.49%) disagree ratings, 141 repetition ratings (3.66%), off-topic (3.35%). So the majority of participants has used ratings to express agreement/disagreement with posted comments.

606 (out of 954) comments have received at least one agreement rating, while 375 comments received at least one disagree rating. Next, 95 comments received at least one repetition rating, and 28 of them received more than 50% of repetition ratings. 43 comments were rated as off-topic at least once, and 18 comments received more than 50% off-topic ratings. Thus, we can conclude that the number of comments marked as off-topic or repetition was small. We had asked two individuals not involved in the experiment to sort out the comments by hand; a result has been that a significant share of repetitions and off-topic comments actually is detected.

To summarize 729 comments received at least one rating; this represents 76% of all comments. This serves as an indication that our proposed approach was well accepted, and that the level of participation was rather satisfying. Agreement and disagreement ratings have shown that participants have followed our suggestion that ratings represent opinion expression, and we can see that ratings options were used extensively.

Qualitative Assessment of Comments. In order to introduce further qualitative measures for the evaluation of comments we define two notions: *rated* and *relevant* comments. A *rated* comment is one which has received at least one rating. A *relevant* comment is one which has received off-topic/repetition ratings in less than 50% of all its ratings. Comments with no ratings are relevant by definition. We use these notions to examine the potential effects of stricter weighting and scoring schemes. The necessity of such stricter rules has come up when examining comments manually. There are some borderline comments close to being repetitions or featuring an argument with a somewhat loose connection to the discussion topic. Quite a number of them are unrated since a reader might have found these comments too difficult/too tedious to rate. Out of 954 comments, 913 are *relevant* (95.7%), 685 are *rated* and *relevant* (71.8%). The data presented reveals the significant share of rated comments. Thus, weights and scores as defined so far would be meaningful even with the stricter selection of comments.

Next, we have observed a significant correlation between the number of posts per participant and the number of his posts that have been rated ($r = 0.8482$, $p < 0.001$). There is also a significant correlation between the number of posts per participant and the relevance of his posts ($< 50\%$ off-topic/repetition ratings) ($r = 0.9765$, $p < 0.001$). When combining these two measures for rated and at the same time relevant posts, we can also confirm a significant correlation between the number of posts and the one of relevant and rated posts ($r = 0.7936$, $p < 0.001$). In Figure 5.7 we can see the number of participants with a certain share of rated and relevant posts in all their posts. We conclude that participants who have participated in the discussion more actively also received significant attention from the community. In other words, more engaged participants were also better discussants, according to the ratings. Furthermore, 80% of the participants have posted *relevant/rated* comments with 80% of their posts. This speaks in favor of the relevance of the discussion and of a large share of good discussants with meaningful posts.

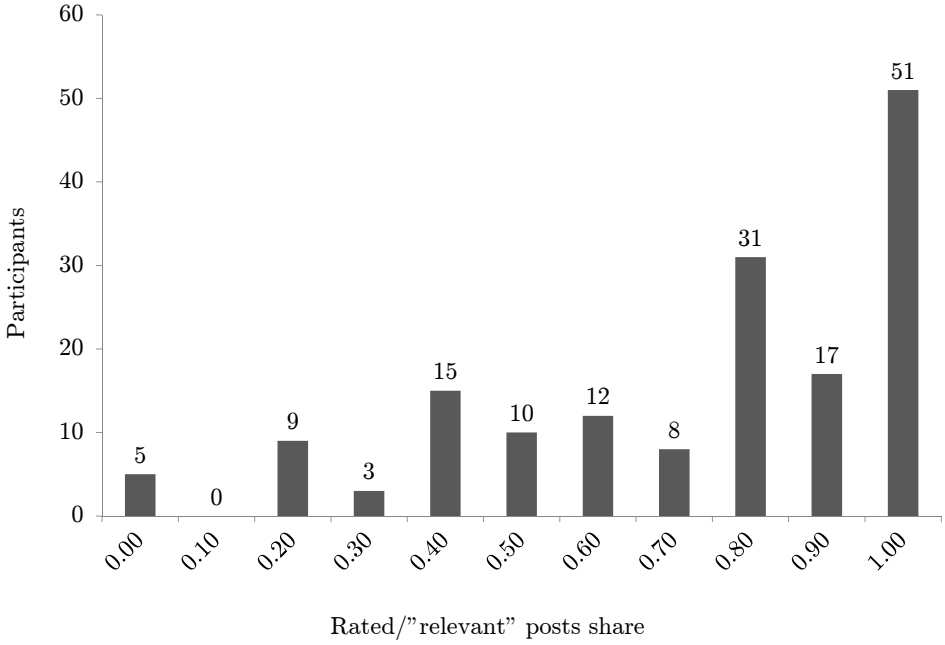


Figure 5.7: Participants per *rated* and *relevant* post share

Additionally, we have counted the numbers of *rated* and *relevant* comments per type. Figure 5.8 shows the distribution of *rated*, *relevant* and *rated/relevant* comments per type.

In the Figure 5.9 graphs the coverage of comments with ratings of different types. E.g., 606 posts have received at least one agreement rating, 375 comments at least one disagreement rating.

Finally, we have compared the comments types specified by authors to feedback information on comment types by other participants by hand. This has not always matched well. In particular, proposal extensions have sometimes been perceived as pro or contra argument comments and vice versa. When studying the robustness of our approach in a later subsection, we will check to which extent these mismatches play a role. Anticipating the respective result, it turns out that the difference is not large. Nevertheless, a potential improvement of our approach could be to allow for such reasonable mismatches, while taking real mismatches (e.g., a fraction of participants stating that a posting is a pro argument and another fraction stating that it is a contra argument) into account. With such a refinement, it would still be plausible why participants should strive to identify the correct type of a comment.

Summary. Participants have been quite active and constant in participating in the discussion. When we look at the participation level and especially the share of participants with a large share of rated and relevant posts, we can conclude that participants have adopted our deliberation approach quite well. The discussion flow has been continuous and lively, responding to the arguments with pro and contra arguments. The majority of

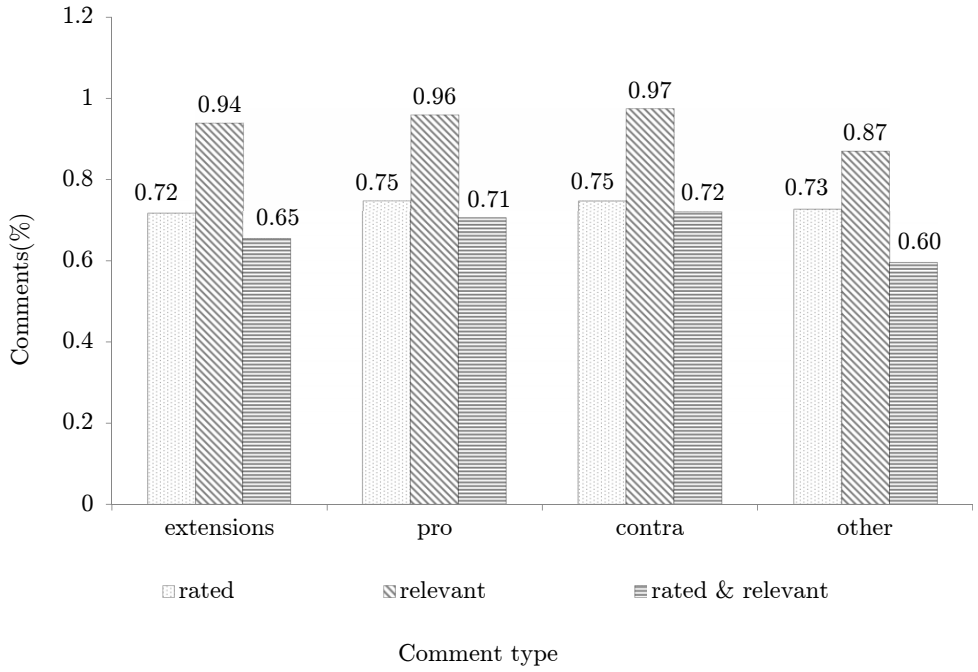


Figure 5.8: Analysis of comments

comments has received ratings; this is good because they are a prerequisite for the evaluation of comments and proposals.

5.7.2 Weighting scheme

Our weighting scheme is based on formal criteria to facilitate efficient discussions without off-topic or repetition comments, offensive and aggressive tone etc.

In Figure 5.10, a set of indicators and the distribution of their values among participants are shown. In the course of the experiment, we have observed problems with our implementation of *hfmscore*, the score assigned by the peer prediction method, so we had omitted it as an argument of the minimum function. We had announced to the participants that we intend to fix it in short time, but ultimately have not been able to do so within the four-week discussion time, mainly because administration issues have required a lot of our attention. As previously stated, indicators such as breadth, engagement, and individuality are normalized to arrive at a uniform distribution of their values, in contrast to indicators that refer to a minority behavior. As expected, there are only few participants who have performed very badly regarding writing style, cf. Figure 5.11. On the other hand, normalized indicators, except for individuality, are grouped in ranges by performance and are less differentiated. Individuality is highly discriminative for participants, and, thus, prevents misuse of the system and herd behavior.

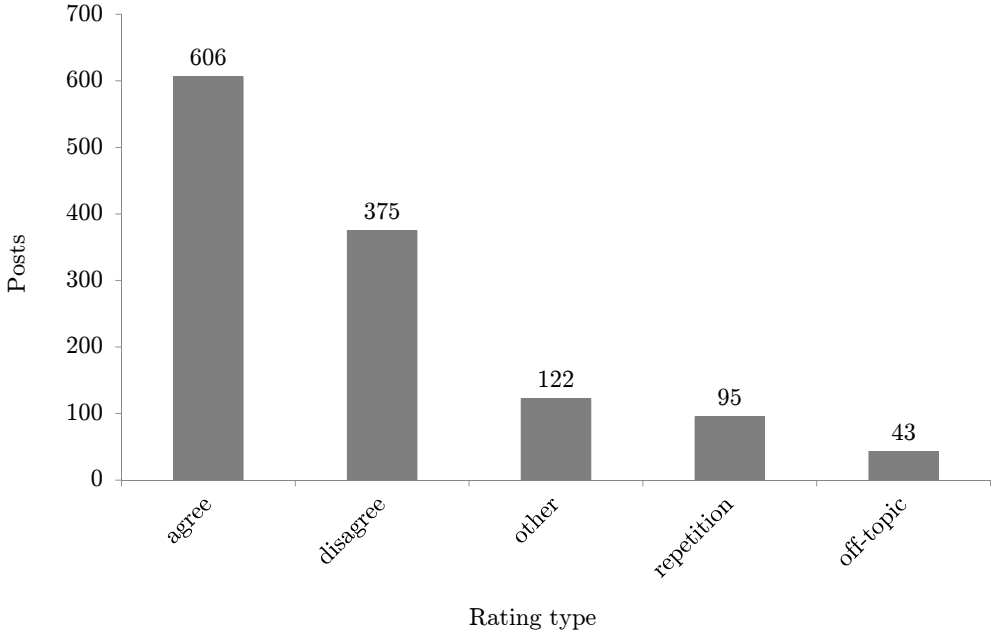


Figure 5.9: Ratings per posts

The fact that individuality is discriminative can also be seen in Figure 5.11. The chart shows the number of participants where the respective indicator value was minimal, compared to all other indicator values of the participant. I.e., when looking at the distribution of the engagement indicator, one can see that 20 participants had performed in the range $[0.9, 1]$ and $[0.8, 0.9]$, the value of 24 participants has been between 0.7 and 0.8 and so on. By redefining indicators or normalizing them differently, the results could be quite different. The rationale behind discussing the concrete indicators values in our setting has been to show the performance of the experiment community and its effect on the evaluation of comments and proposals.

Given our data, we can confirm a significant correlation between the number of comments posted by a participant and his weight ($r = 0.6783$, $p < 0.0005$). I.e., an engaged participants with broad interests and cooperative behavior should be recognized as a good discussant.

5.7.3 Comment scoring scheme

Comments are evaluated based on the respective scoring scheme. We clearly expect that comments that have received many agreement ratings are likely to be rated higher that the ones with only a few agreements or even a lot of disagreement ratings. The respective correlation is $r = 0.46265$, $p < 0.0005$. We also expect comments rated or posted by high-weight participants to score higher than in case of low-weight raters or authors. The correlation between comment scores and author weights is $p = 0.4530$, $p < 0.0005$. The correlation between comment scores and rater weights is lower, with $r = 0.2625$, but it

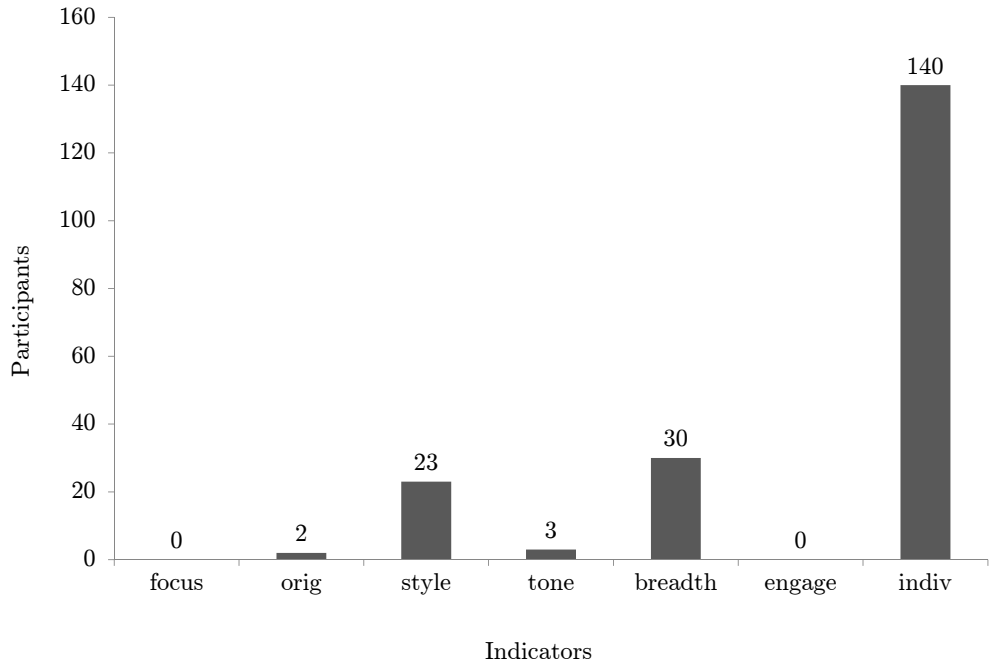


Figure 5.10: Participants per indicator values

is confirmed with $p < 0.0005$ due to the sample size. Furthermore, posters with higher weights have received more ratings for their comments, and the correlation is significant, with $r = 0.4826$, $p < 0.001$. When we compare the weights of posters and the sum or average of their comment scores, we see significant correlations, $r = 0.5771$, 0.4149 respectively with $p < 0.001$. Based on this, we conclude that discussants with higher weights post comments rated with higher scores. The average score of the comments of these participants was higher, and they received more ratings for their comments. Thus, the community has perceived higher-weighted participants as discussants with more interesting contributions.

The listed results indicate that the comment scores reflect the criteria that we deem important in the evaluation, such as number of ratings received, the reputation of the author and the raters as well as agreement status in the community.

Figure 5.12 shows the average scores of best posts by post date.

Best posts are the top quarter of all posts based on score. When analyzing the connection between the posting date and the final score, we can see that, as expected, early posts in general receive more attention and thus, the scores are higher. Still, if we look at the distribution of the best posts (the top quarter of the scores of all posts), there are peaks related to forum participation, i.e., the opening of the forum discussion and the point of time when the discussion has been extended. Latter comments can also achieve high scores. This speaks in favor of the scoring scheme.

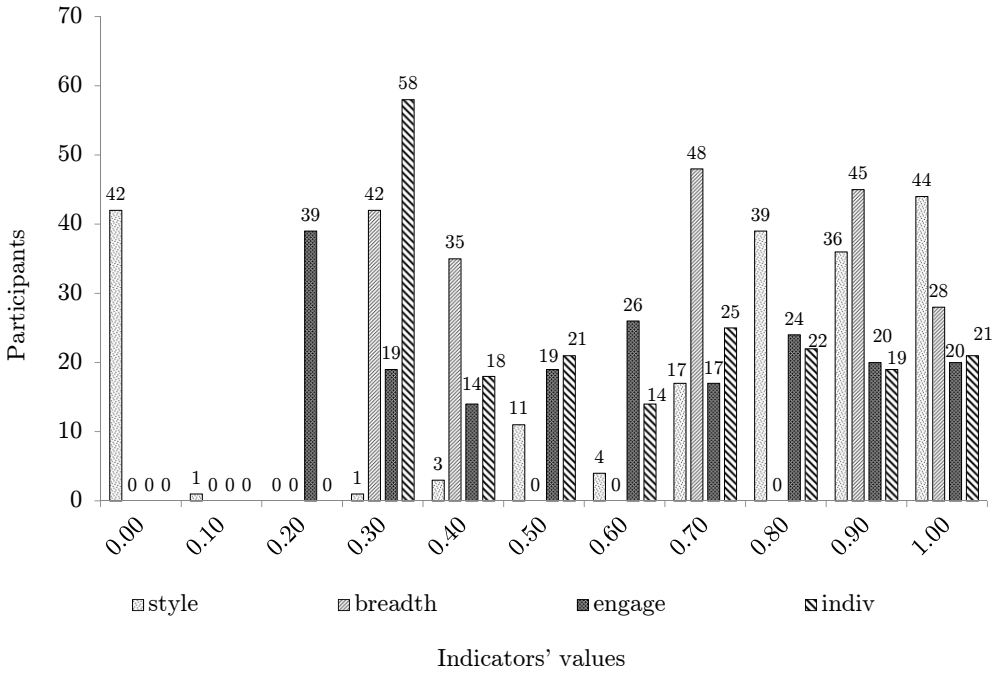


Figure 5.11: Participants per minimal indicator value

5.7.4 Proposal scoring scheme

To analyze the effects of the proposal scoring scheme, we look at the outcome of proposal ratings and their common characteristics. An objective truth criterion does not exist in this case, and this represents a big challenge for assessing the scoring scheme. There are some rather obvious results, e.g., there is a significant correlation between the proposal score and the number of pro arguments, $r = 0.5899$, $p < 0.001$. Our analysis has not confirmed a significant correlation between the score of the poster and the number of positive proposals, e.g., proposals that have received more pro arguments than contra arguments. The highest number of positive proposals posted by one poster was three. The weight of this participant is in the best fifth of the ones of all participants. Six other participants posted two proposals each, and their weights vary from 0.402 in the top 45% of the participants to 0.8826, the second best participant. Although we cannot directly connect participant weights and proposal scores, we see some indications that participants who performed better regarding our formal criteria posted comparably better proposals. These indications include the correlation between author weights and comment scores or even between average comment scores and author weights.

In order to confirm the robustness of our approach in terms of its sensitivity to changes and different evaluation criteria, we have tried out several alternative scoring schemes and have examined their effect on the outcome (proposal scores in particular). We deem this necessary, since we had observed some imperfections of our approach in the course of our experiment, as mentioned earlier. In particular, there are mismatches of comment types

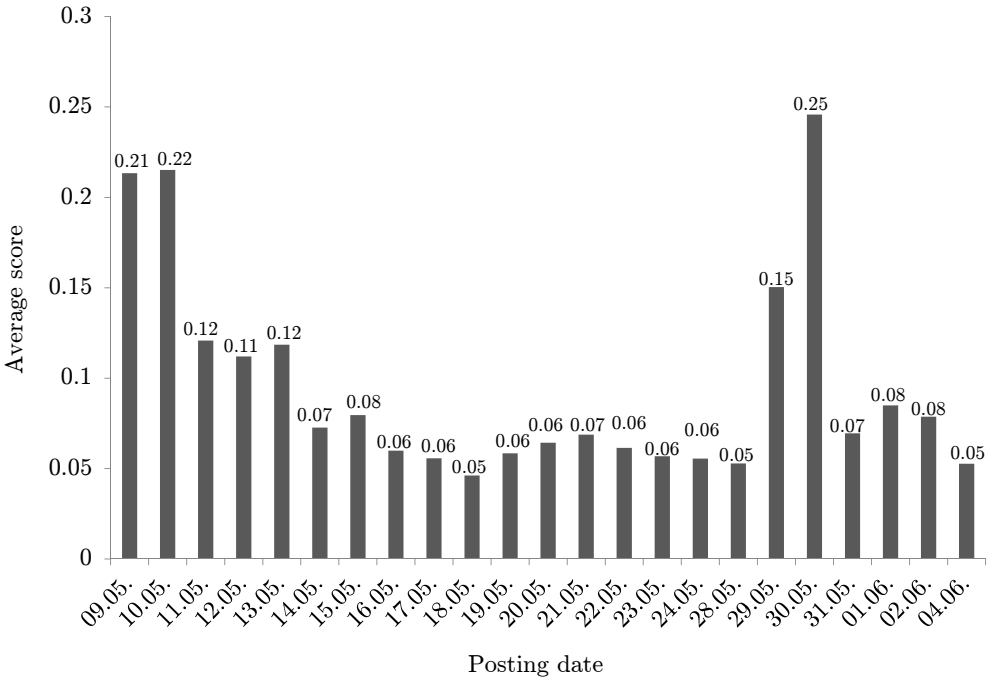


Figure 5.12: Best scores averages distribution per post date

by authors and other participants as well as unrated and irrelevant comments. We have examined how slight modifications of the approach affect the outcome based on the data collected in this current experiment. Ideally, we can verify that these slight changes do not influence the output significantly. We have studied the following alternatives:

(1) Use simple count of agreements and disagreements as up and down votes instead of using weights to calculate comment scores. These calculated comment scores are then used as input of the proposal scoring scheme.

(2) Stricter rules for useful comments, namely, comments that are not rated or not relevant are ignored.

(3) Resolve comment types based on type ratings with 60% and 80% certainty. In our current proposal-scoring scheme, we take the argument type as specified by the author at face value. To include comment-type checking e.g., if a comment is a pro argument or a contra argument, we have taken the type ratings received for comments into account. Here, we propose a threshold of 60% (80%) certainty when resolving the type. Comments whose type is not verified are ignored.

5.8 Conclusion

We described and evaluated a new approach to facilitate efficient online deliberation. We have proposed a novel approach for identifying and evaluation of problem solutions in online

settings, based on the discussion itself and its structure. In order to achieve a discussion structure that gives way to an evaluation of solution options we have come up with various extensions of conventional forums and discussion structures. Our primary aim has been to organize the discussion around the deliberative principle of carefully considering pro and contra points. The essence of our approach is a three-step evaluation: First, based on a set of formal criteria, we rate users and assign them weights. Next, comments are evaluated based on the agreement/disagreement ratings of the argumentation referring to them and the weights of raters and authors. Finally, solutions of discussion issues are evaluated based on the scores of pro and contra arguments posted.

We have evaluated our approach by conducting an experiment and a subsequent survey with individuals discussing topics of relevance for this particular community. The result of the questionnaire confirms that the community had been satisfied with our forum model and the respective decisions. In the next step, we have systematically analyzed the data collected in an experiment with around 200 participants. We can verify active participation during the four weeks of forum discussion. Significant numbers of comments and ratings have been generated, and a large share of participants has been active. We conclude that our proposed approach for online deliberation was well accepted in the community. In particular, the analysis of participant weights and of comment and proposal scores has demonstrated compliance with our assessment criteria such as community consensus, observance of the deliberation principle and fulfillment of requirements for efficient discussions. In our specific setting, the approach has shown to be very effective when dealing with repetitions, off-topic comments, or aggressive tone. We see this as an indication that it might perform similarly well in other settings.

6 Formal Model

Discussion forums play an important role when dealing with issues of common interests. 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 mainly orthogonally to these dimensions, it is of course interesting to know whether a community agrees or disagrees with a certain comment. One-dimensional feedback options such as 'thumbs up', 'like' buttons etc. are too undifferentiated to this end. Feedback where different types are feasible in turn, i.e., participants can assess a post by someone else according to different criteria, seems to be a viable solution to this problem.

Participants in discussions have different interests, which motivate them to issue feedback on posts. In the presence of feedback of different types, they can behave strategically in order to back up their specific interests and to push their opinion, cf. [6].

Example 1: Let us 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, we observe a case 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 chapter, we present our study that deals with this specific setting, namely *various rating strategies* in the presence of *feedback of different types* [128]. 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: We take two Comments A and B posted in the discussion forum . 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.

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

This example has several important implications: P is able to gain the advantage because the feedback by others (or the fact that no one else has given feedback yet in this specific case) 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. More specifically, we study a setting where ratings are not published at all until the discussion, the time window when issuing feedback is possible, ends. Such a setting also is 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 an equilibrium 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 [129].

6.1 Challenges

The problem investigated here is challenging, for several reasons: First, building a formal model of forum discussions is quite challenging. Models of social systems are complex, with imprecise, incomplete and inconsistent theories [100]. While it should be sufficiently exhaustive and representative to allow for meaningful conclusions; a model must not be overly complex. Our objective was to mimic the discussion structure of a forum, 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 particular context. In our previous work based on user experiments [125], we have followed a two-step evaluation using a weighting and a 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, higher degrees of 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 an individual 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, in order 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.

6.2 Design

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 [101]. 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 for 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 formal model features posting comments or ratings of comments authored by other participants. A participant can post a rating of the following values: Agree, Disagree, Repetition. A participant does not have to rate a certain comment at all.

To illustrate the nature of the game, we use Figure 6.1. 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

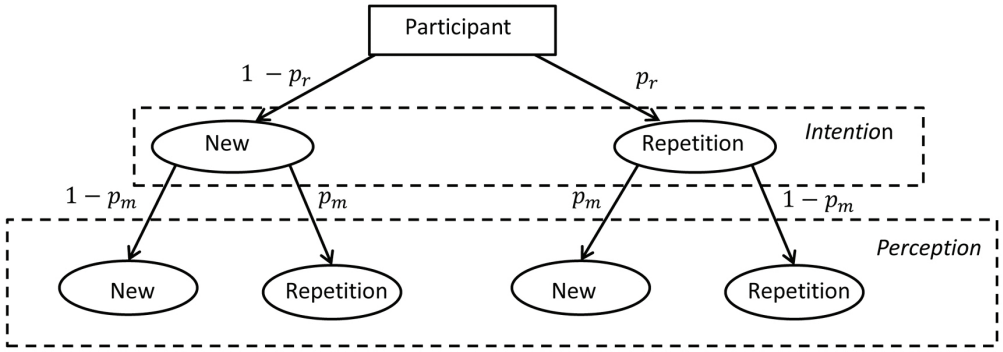


Figure 6.1: Nature of the game

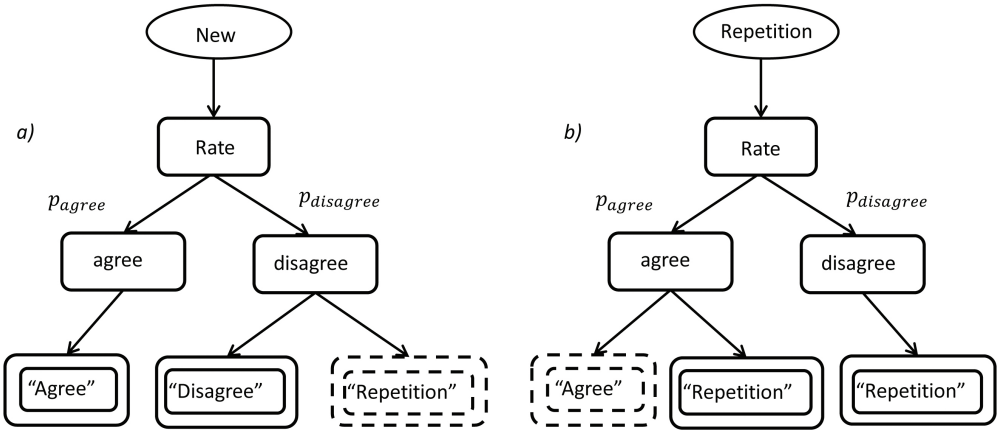


Figure 6.2: Participants moves for new/repetition comments

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

Figure 6.2 a) shows the possible moves of a participant when perceiving a comment as new, Figure 6.2 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 6.2 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.

6.2.1 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 6.3.3. 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 the majority of rates whether a comment is a repetition, this indicator is high. The rationale behind this criterion is to try to assess whether participants generate ratings truthfully. Note that this criterion only takes the "Repetition" or "Agree"/"Disagree" ratings. 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's 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.

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 and 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 with $R^{subject}_{aname}(r.comment)$. $R^{create}_{repetition}(j)$ is the set of all "Repetition" ratings Participant j issued. Whereas $R^{subject}(r.comment)$ is the set of all ratings of the comment. First, we calculate a value for each issued rating of 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 all ratings' scores per value, i.e. "Repetition" or "Agree"/"Disagree". 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.

6.2.2 Scoring Scheme

A scoring scheme evaluates comments based on the ratings received and, in our case, the respective weight of raters and authors. Recall that the rationale behind feedback of different types is better quality control, i.e., to be able to 'sort out' certain unwanted comments. In our case, these are repetitions of previous posts. Thus, 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:

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

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, whereas $weight(issuer(r))$ 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, whereas $R^{subject}_{opinion:agree}(k)$ 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 Comment 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, namely, but not necessarily in the form of positive feedback throughout.

6.3 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 against different rating strategies. Namely, we want to check if 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 at least the same or higher utility from one of the strategies.

In more detail, as feedback by others is hidden from participants for the time being, 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 at least the same or higher utility in all classes. As a second step, 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 misperception of a comment (p_m), i.e., rater perception of a comment differs from the original intention of the author (see Figure 6.1). Perceiving a new comment as repetition and repetition as new can happen with certain probabilities. For the sake of keeping 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) Probabilities that determine the rating behavior of synthetic participants (p_{ur}), i.e., whether a synthetic participant rates a comment that he perceives as new not as a repetition and vice versa (see Figure 6.2). Note that this parameter is only relevant when studying trembling: When answering the question whether 'always truthful' is a probabilistic equilibrium, these probabilities are set to zero. We only vary them when 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 the agreement/disagreement ratio in the community, 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.

6.3.1 Implementation

The formal model we have proposed has the following elements:

- Participants
- Comments
- Ratings
- 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 6.2. 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, except for the respective authors (or they have chosen not to give feedback regarding some comments). 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, in our context, 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 respective participant regarding the comment. The following moves are available:

- Write a comment (w)
- Rate a comment with ratings "Agree" (a), "Disagree" (d), "Repetition" (r)
- 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 row, i.e., exactly one author per comment.

$$\begin{bmatrix}(1,\,0,\,0,\,0) & (0,\,1,\,0,\,0) & (1,\,0,\,0,\,0)\\ (0,\,0,\,1,\,0) & (1,\,0,\,0,\,0) & (0,\,0,\,0,\,1)\end{bmatrix} \tag{6.1}$$

The action matrix for two participants ($P1, P2$) and three comments ($K1, K2, K3$) serves as an illustration (see 6.1). 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 $K1, K3$, respectively. Obviously, $K2$ is the only comment that has not received any rating yet.

6.3.2 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 certain strategies 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 well as 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.

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

(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].

6.3.3 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 for 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, possibly in order to create an incentive for participation. 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 $utility\_rep(v, s, s')$

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

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.

$$\begin{aligned} utility\_count(v, s, s') = \\ \big|\{k \in K^+ : score(k, s) \neq null \wedge score(k, s') \neq null \\ \wedge score(k, s) \geq score(k, s')+ \\ \big|\{k \in K^- : score(k, s) \neq null \\ \wedge score(k, s') \neq null \wedge score(k, s) \leq score(k, s') \end{aligned} \quad (6.2)$$

(2) The utility function *utility_sum* quantifies the benefit of the controlled participant based on the scores of comments he agrees and disagrees with.

$$\begin{aligned} utility\_count(v, s) = \\ \sum_{k \in \{k \in K^+ : score(k,s) \neq null\}} score(k, s) - \\ \sum_{k \in \{k \in K^- : score(k,s) \neq null\}} score(k, s) \end{aligned} \quad (6.3)$$

(3) The relative utility function *utility_rep* has the valuation function v, the observed strategy and the reference strategy as arguments.

$$\begin{aligned} utility\_rep(v, s, s') = \\ \big|k \in K^+ \wedge score(k, s) \neq null \wedge score(k, s') = null\big| + \\ \big|k \in K^- \wedge score(k, s) = null \wedge score(k, s') \neq null\big| \end{aligned} \quad (6.4)$$

(4) This utility function takes the weight of the controlled participant as success criterion. $weight(p\_controlled, s)$, respectively is another criterion we used for quantification of the controlled participant's benefit when using strategy s.

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

6.3.4 Simulations

We initialize the formal model with certain number of participants and certain number of comments (*num_participants*, *num_comments*), respectively. Comments are completely randomly authored by the participants. In the next step, we generate equivalence classes for the comments. The details on equivalence classes will follow. We generate a number of ratings, e.g. we set the maximum number of ratings and the final number of generated ratings highly depends 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 at least the same or 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 6.1 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 6.1: Simulation setup parameters

6.4 Results

A first important insight from our simulations with the default setting is that 'always untruthful' strategy 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' strategy brings at least the same or higher utility than the other strategy.

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' strategy 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 any 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 presented in Table 6.2. 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.

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

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 6.2: Results when varying the share of untruthful ratings

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 at least the same or greater utility. We see this as a positive result; even in the case of mixed strategies, untruthful behavior does not pay off.

6.5 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 chapter, we have presented the results of our study which examines the effects 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.

7 Conclusions

In the thesis, we have presented our work conducted during the project which lasted for three years. The main outcome of our work is a novel approach for organizing constructive discussions based on arguments and reasoning, reducing off-topic and redundant contributions and insisting on clarity and balanced, polite considerations. We have proposed a comprehensive forum model that incorporates simple argumentation model to bring the structure in the discussion categorizing comments as pro or contra arguments. Moreover, we have suggested different types of feedback to sort out comments e.g. recognize redundant or repeated considerations. Finally, the forum model employs weighting scheme and scoring scheme to assess both participants and contributions.

Our motivation to take this particular research direction stems from increasing number of online discussion forums, groups and projects and their growing importance in organizing public life. Accordingly, the spectrum of respective communities is broad in terms of different fields (science, technology) and size of communities from municipalities to large administrative units. Regardless of high variety of settings, topics and levels of consideration they all face common issues such as: growing amount of content, redundancy, divergence from the discussion topic, incompleteness, aggressive and offensive behavior. Consequently, the problem of efficient assessment of contributions arises.

Our main research question was whether it would be possible to conduct evaluation based on the discussion itself and its structure. Furthermore, the question is how to uniformly and comprehensively evaluate discussion points that are topic specific, and opinion based and finally, extract important, valuable ideas, arguments as the main goal of the evaluation process.

7.1 Our approach

The forum model, we have proposed, aims to assist group-decision making, e.g. proposing and selecting solutions to issues discussed in online settings. It relies on relatively simple argumentation model and feedback of different types to facilitate evaluation. The process of evaluation itself consists of three steps:

(1) assigning weights to participants based on the set of formal criteria such as degree of engagement in the discussion to stimulate desirable behavior , e.g., originality of arguments, focus on the topic in question etc.;

(2) assigning scores to comments, considering the weights of authors and raters and the agreements/disagreements of the community with the expressed argumentation;

(3) assigning scores to proposals, based on the scores of the pro and contra arguments.

Hence, it is important to verify to which extent individuals have understood and accepted the approach. In the next step, we were interested in how model had performed to identify 'good' discussants, strong arguments and valuable proposals. Finally, another focus of our work was to study the robustness of the approach with regard to minor changes.

7.1.1 Assessment and results

To evaluate our proposed approach we have decided for the experimental method. For this purpose, we have initiated online discussion in a group of 100 participants on topics of interests in the community. Subsequently, participants completed a survey to share their experience in the discussion based on our proposed forum model. In the next step, we have conducted a comprehensive analysis of the data collected during the experiment. Thus, we tested a set of hypotheses based on participants' perception of our forum model along with factual data shown in calculated indicators, scores and charts.

The questionnaire results have verified that:

Participants have deemed our weighting scheme fair.

Participants have confirmed tight connection between perceived usefulness of the decision-making scheme and the perceived fairness of the weighting model.

The evaluation of the decision-making scheme and the perceived quality of the decisions are correlated according to the participants' reports.

The perceived quality of the decisions is positively correlated with the participants' feeling that their opinion is respected.

Participants were honest when rating contributions of others according to the results in both basic and control question.

Beyond the questionnaire results, analysis of the data collected during the experiment aimed at a broader perspective in terms of performance of our approach and its robustness even in the absence of an objective truth criterion. We examined the following indications:

We looked at the posting behavior of participants in terms of number of comments and ratings, their distribution along with the relation to the assessment of comments. Participation level and especially the share of participants with a large share of rated and relevant (no off-topic or repeated) posts confirm that participants have adopted our deliberation approach quite well. The discussion flow has been continuous and lively, responding to the arguments with pro and contra arguments. The majority of comments have received ratings. Therefore, participants have confirmed the intrinsic motivation to contribute.

Data analysis related to weighting scheme relies on formal criteria indicators, their individual values, distribution and the degree of discrimination in the community. Indicator of individuality has been recognized as more discriminative in the community to prevents misuse of the system and herd behavior. Moreover, participants adherence to criteria has lowered dramatically posts with unclear writing style or provocative, aggressive tone.

Comment scoring scheme analysis refers to comments' scores, their distribution and correlation with authors' weight. Analysis shows that discussants with higher weights posted

comments rated with higher scores. The average score of the comments of these participants was higher, and they received more ratings for their comments. Thus, the community has perceived higher-weighted participants as discussants with more interesting contributions.

Proposal scoring scheme analysis addresses effects of the schema on scores of different proposals and their common characteristics. Although we cannot directly connect participant weights and proposal scores, we see some indications that participants who performed better regarding our formal criteria posted comparably better proposals.

Finally, in order to confirm the robustness of our approach in terms of its sensitivity to changes and different evaluation criteria, we have tried out several alternative scoring schemes such as simple count of agreements and disagreements, more restrictive rules for considering an argument validity or type (pro or contra). With no doubts, we have confirmed robustness of our approach in terms of its effects on the outcome (especially proposal scores).

We conclude that our proposed approach for online deliberation was well accepted in the community. In particular, the analysis of participants' weights and comments' and proposals' scores has demonstrated compliance with our assessment criteria such as community consensus, respect of the deliberation principle and fulfillment of requirements for constructive discussions to increase quality and clarity of posts. In our specific settings, the approach has shown to be very effective when dealing with repetitions, off-topic comments, or aggressive tone. We see this as an indication that it might perform similarly well in other settings.

Proceeding with an experiment as an evaluation method was supported by a good opportunity to employee the university students attending our course offering them to discuss the topics of their interests. This has given a way for trying out our specific approach in a real situation offering lively and authentic feedback. Still, the experiment in total with the analysis of survey results and data collected has lasted for more than a year. Hence, organizing further experiments appeared to be too costly in terms of time and financially, thus, we have decided to use formal analysis to carry on further assessment of our approach.

Formal models and game-theoretic approach have appeared to be very handful in modeling self-interested multi-agents interactions. Similarly, in online discussions participants have different interests and opinions which motivate them to pick different strategies to back up their specific interests and to push their opinion. In the context of different feedback options as foundation of the evaluation system, inquiring how different strategies can influence outcome is very important. Next, studying whether truthful behavior is an equilibrium strategy is meaningful, i.e., is it expected that a truthful behavior would be the most beneficial , assuming that everybody else acts truthfully. Finally, robustness of our approach to accommodate various rating strategy has been an important focus in our study.

For this purpose, we have developed a formal model that relies on game-theoretic approach focused on examining one type of unwanted behavior, namely, posting repetitions. 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 rating strategies.

As a main result it has been confirmed that our forum model gives preferences to truthful rating strategy, e.g. the strategy 'always truthful' is an equilibrium strategy. It also is

superior to the reference strategy with some modifications of the model. Although the formal model has incorporated particular scoring schemes and utility functions, our approach is sufficiently modular to take other schemes or other utility functions instead.

7.2 Future work

Our work aims at addressing common difficulties online discussions face in many applications. The work itself presents comprehensive consideration of a complex problem. Evaluation of our approach has included experimental method, backed up with survey results. Additionally, we have employed formal analysis to gain further insights and verifications of the proposed forum model. Nevertheless, we have only scratched the surface of a very complex problem of modern social networks, dynamics of social interactions, motivation to contribute, behavioral strategies, assessment of content generated by many to extract valuable points, conclusions. Therefore, in the context of our particular project there are many possibilities to continue the work.

Our proposed approach has shown to be very promising in addressing the problem of organizing constructive online discussion based on argumentation and reasoning and evaluating contributions based on the discussion itself and its structure. Further experiments with some other communities and topics could be also very useful in gaining additional insights how different communities react to our proposed forum model. In following experiments we could also vary some settings in order to improve performances or address different communities and their specifics. To conduct an experiment is a long process from planning till execution phase, collecting and analysis of results. A choice of respective community, topics of interests and incentives can influence greatly the outcomes.

Our work could also be applied in other settings than discussion forums such as management of ideas. This was one of possible research directions that we have also considered. The idea has inspired some endeavors of ours to organize an experimental study with one of the leading company in chemical industry. Unfortunately, due to some organizational issues, project was never set in motion.

Regarding the formal analysis, opportunities for further consideration are numerous. We have picked one special setting that we have deemed the most important and critical in our context. Since our entire evaluation system is highly relying on participants' feedback, we were interested in examining how beneficial would be for a participant to behave untruthfully when other behaves truthfully. Formal analysis can be extended to some other aspects of our formal model either participants' behavior such as rating or posting behavior or varying characteristics of settings such as quality of comments (off-topic or redundant comments and valid arguments), ratio of pro or contra comments etc. Finally, examining different behavioral strategies or parameters of the settings and their influence on the evaluation could be interesting new undertakings to further examine the strengths and weaknesses of our proposed model.

Appendix

.1 Questionnaire - Discussion forum ”Database systems”

1. Have you read the technical documentation that describes used criteria for participants' weighting and scoring of proposals? (1 = not at all, 2 = took a glance at it, 3 = read it, 4 = read it carefully, 5 = read it carefully and understood it)

   1 2 3 4 5

2. How many comments have you posted approximately?

3. In your opinion, is it fair to use these formal criteria to weight opinions in contrast to having all the opinions equally weighted? (1 = not at all, . . . 5 = yes, indeed)

   1 2 3 4 5

4. What is your opinion on the following formal criteria:

   - Focus (1= impractical, . . . , 5=very useful) 1 2 3 4 5
   - Originality (1= impractical, . . . , 5=very useful) 1 2 3 4 5
   - Style (1= impractical, . . . , 5=very useful) 1 2 3 4 5
   - Tone (1= impractical, . . . , 5=very useful) 1 2 3 4 5
   - Individuality (1= impractical, . . . , 5=very useful) 1 2 3 4 5
   - Engagement (1= impractical, . . . , 5=very useful) 1 2 3 4 5
   - Breadth (1= impractical, . . . , 5=very useful) 1 2 3 4 5
   - Honesty (1= impractical, . . . , 5=very useful) 1 2 3 4 5

5. Have these formal criteria affected your behavior when posting comments or ratings? (1=not at all, . . . , 5=very strong)

   1 2 3 4 5

6. If yes, please explain, how was your behavior affected?

7. We internally have calculated a value for each criterion and each participant and derived a weight based on these values and published it. Have you been influenced by your weight when posting comments or ratings? (1=not at all, ..., 5=very strong)

   1 2 3 4 5

8. Did you have the impression that you could influence your weight? (1=not at all, ..., 5=yes, very strong)

   1 2 3 4 5

9. Can you please state in your own words your understanding of the criteria used for proposals' scoring?

10. What do you think about our decision-making scheme? (1= impractical, ..., 5=very useful)

    1 2 3 4 5

11. If you find our decision-making scheme not acceptable, please state why?

12. How often have you expressed your opinion truthfully when posting comments? (1=never, ..., 5=always)

    1 2 3 4 5

13. How often have you expressed your opinion truthfully when posting ratings? (1=never, ..., 5=always)

    1 2 3 4 5

14. If you have not been honest, please explain your motivation to do so?

15. According to your estimation how many participants in this study (in %) have always been honest? %

16. What do you think of our approach to evaluate proposals in comparison to simple poll voting, considering following criteria? (1=poor, ..., 5=very good)

    a) Appropriateness of the decisions 1 2 3 4 5

    b) Forum user-friendliness 1 2 3 4 5

    c) Respect of the opinion 1 2 3 4 5

    d) Option to define proposal extensions 1 2 3 4 5

17. Please compare the discussion in our forum (different types of comments and multi–facet ratings) with the discussion in conventional forums. Do you like our forum model more? (1=not at all, . . . , 5=yes, indeed)

    1 2 3 4 5

18. If you do not like our forum model, can you please state why is it so?

19. Considering the final decisions, do you find them more acceptable in comparison to take the decisions by simple poll voting? (1= not at all, . . . , 5 = yes, indeed)

    1 2 3 4 5

20. Do you have any suggestion how we could improve our forum model from conceptual or technical point of view?

21. Supposing that we use the same forum next year to generate the data for SQL exercises, what would be your suggestion for new discussion issues? Our budget for each decision is approx. EUR 500.

Bibliography

[1] 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.

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

[3] Mark Klein, "Enabling large-scale deliberation using attention-mediation metrics," vol., no., 2012. [Online]. Available: http://dx.doi.org/10.1007/s10606-012-9156-4

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

[5] T. Kriplean and J. Morgan and D. Freelon and A. Borning and L. Bennett, "Supporting Reflective Public Thought with ConsiderIt," in *CSCW*,

[6] Arpita Ghosh and Patrick Hummel, "Implementing optimal outcomes in social computing: A game-theoretic approach," in *Proceedings of the 21st World Wide Web Conference 2012, WWW 2012, Lyon, France, April 16-20, 2012*, [Online]. Available: http://doi.acm.org/10.1145/2187836.2187910

[7] Hans K. Klein, "Tocqueville in cyberspace: Using the internet for citizen associations," vol., no., 1999. [Online]. Available: http://dx.doi.org/10.1080/019722499128376

[8] Raymond J. Pingree, *A New Approch to Three Problems in Deliberation, in Online Deliberation: Design, Research, and Practice.* Stanford, CA: Stanford University, 2009.

[9] Simone Chambers, *Reasonable Democracy.* NY: Ithaka, 1996.

[10] J. Cohen, ch., Deliberation and Democratic Legitimacy.

[11] Carpini, Michael X Delli and Cook, Fay Lomax and Jacobs, Lawrence R, "Public deliberation, discursive participation, and citizen engagement: A review of the empirical literature," vol., 2004.

[12] ch., Deliberation as Discussion.

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

[14] ——, *The Voice of the People.* Binghamton, NY: Vail-Ballou Press, 1995.

[15] J. Gastil, *By Popular Demand: Revitalizing Representative Democracy Through Deliberative Elections.* Oakland, CA: University of California Press, 2000.

[16] A. Gutmann, and D. Thompson, *Democracy and disagreement.* Harvard University Press, 2009.

[17] Converse, Philip E, "The nature of belief systems in mass publics," in *Collection*,

[18] Carpini, Michael X Delli and Keeter, Scott, *What Americans know about politics and why it matters*. Yale University Press, 1996.

[19] Kuklinski,James H. and Quirk,Paul J. and Jerit,Jennifer and Schwieder,David and Rich,Robert F., "Misinformation and the Currency of Democratic Citizenship," vol., 8 2000. [Online]. Available: http://journals.cambridge.org/article_S0022381600000335

[20] Feldman, Stanley, "Measuring issue preferences: The problem of response instability," vol., no., 1989.

[21] Rosenstone, S.J. and Hansen, J.M. and Reeves, K., *Mobilization, Participation, and Democracy in America*, ser. Longman Classics Edition. Longman, 2003. [Online]. Available: http://books.google.de/books?id=QicLAAAACAAJ

[22] Verba, S. and Schlozman, K.L. and Brady, H.E., *Voice and Equality: Civic Voluntarism in American Politics*. Harvard University Press, 1995. [Online]. Available: http://books.google.it/books?id=YFiCO5f0BKAC

[23] Benhabib, S., "Toward a Deliberative Model of Democratic Legitimacy," in *Democracy and Difference: Contesting The Boundaries of The Political*.

[24] Young, I. M., "Communication and the Other: Beyond Deliberative Democracy," in *Democracy and Difference: Contesting The Boundaries of The Political*.

[25] Sanders, Lynn M, "Against deliberation," 1997.

[26] Jayasena, G. N., & Karunaratna, D.D, "Towards an enabling framework for e-democracy through the integration of citizen participation along the spatial dimension using free and open source technologies," in *Proceedings of the 24th South East Asia Regional Computer Conference, 2007, Bangkok, Thailand, November 18-19, 2007*, 2007.

[27] Mbarika, V.W.A., *Africa's Least Developed Countries' Teledensity Problems and Strategies: Telecommunications Stakeholders Speak*. ME & AGWECAMS Publishers, 2001. [Online]. Available: http://books.google.de/books?id=Dy24AAAAIAAJ

[28] Riley, C. G., "The Changing Role of the Citizen in the E-Governance & E-Democracy Equation," *Commonwealth Centre for e-Governancey*, 2003.

[29] Zahid Parvez, "E-democracy from the perspective of local elected members," vol., no., 2008. [Online]. Available: http://dx.doi.org/10.4018/jegr.2008070102

[30] Funilkul, Suree and Chutimaskul, Wichian, "The Framework for Sustainable E-Democracy Development," vol., no., 2009.

[31] Abinwi C. Nchise, "The Trend of E-Democracy Research: Summary Evidence and Implications," in *13th Annual International Conference on Digital Government Research, dg.o 2012, College Park, MD, USA, June 4-7, 2012*, [Online]. Available: http://doi.acm.org/10.1145/2307729.2307756

[32] Van Dijk, Jan, "Models of democracy and concepts of communication," 2000.

[33] Mutz,Diana C., "Facilitating communication across lines of political difference: The role of mass media," vol., 3 2001. [Online]. Available: http://journals.cambridge.org/article_S0003055401000223

[34] Conover,Pamela Johnson and Searting,Donald D. and Crewe,Ivor M., "The Deliberative Potential of Political Discussion," vol., 1 2002. [Online]. Available: http://journals.cambridge.org/article_S0007123402000029

[35] Gastil, John, *Political Communication and Deliberation*. SAGE Publications, Inc, 2008. [Online]. Available: http://books.google.de/books?id=Dy24AAAAIAAJ

[36] Edwards, Arthur and Scott Wright, "Moderation in Government-Run Online Fora," in *Encyclopedia of Information Science and Technology*.

[37] Shanto Iyengar, Robert C. Luskin and James Fishkin, "Deliberative Public Opinion in Presidential Primaries: Evidence from the Online Deliberative Poll *," in *Voice and Citizenship: Re-thinking Theory and Practice in Political Communication, University of Washington, April 23-24, 2004*, 2004.

[38] Jankowski, Nicholas W and Van Os, Renée, "Internet-based Political Discourse: A Case Study of Electronic Democracy in Hoogeveen," 2004.

[39] Muhlberger, Peter, "The Virtual Agora Project: A Research Design for Studying Democratic Deliberation," vol., 1, 2005.

[40] Sproull, Lee and Kiesler, Sara, "Reducing Social Context Cues: Electronic Mail in Organizational Communication," vol., no., 1986.

[41] Muhlberger, Peter, "Beyond Political Interest: Political Internalization in Political Participation," in *Annual Meeting of the American Political Science Association 2003, Philadelphia, PA, 2003*, 2003.

[42] ——, "Human Agency and the Revitalization of the Public Sphere," vol., 2005.

[43] Douglas Schuler, "Online Civic Deliberation with E-liberate," in *Online Deliberation: Design, Research, & Practice*.

[44] Walton, Douglas, "Justification of Argumentation Schemes," vol., 2005.

[45] Chris Reed and Douglas Walton, "Towards a formal and implemented model of argumentation schemes in agent communication," vol., no., 2005. [Online]. Available: http://dx.doi.org/10.1007/s10458-005-1729-x

[46] Chaim Perelman and L. Olbrechts-Tyteca, *The New Rhetoric: A Treatise on Argumentation*. Notre Dame, 1969.

[47] Hastings, Arthur C., "A Reformulation of the Modes of Reasoning in Argumentation," Ph.D. dissertation, Evanston, Illinois, 1963.

[48] S. E. Toulmin, *The Uses of Argument*, 1958.

[49] Verheij, Bart, "Evaluating Arguments Based on Toulmin's Scheme," vol., no., 3, 2005.

[50] Bart Verheij, "Deflog: on the logical interpretation of prima facie justified assumptions," vol., no., 2003. [Online]. Available: http://dx.doi.org/10.1093/logcom/13.3.319

[51] ——, "Artificial argument assistants for defeasible argumentation," vol., no., 2003. [Online]. Available: http://dx.doi.org/10.1016/S0004-3702(03)00107-3

[52] ——, "Argue! - an implemented system for computer-mediated defeasible argumentation," in *THE TENTH NETHERLANDS/BELGIUM CONFERENCE ON ARTIFICIAL INTELLIGENCE (NAIC '98*,

[53] Douglas Walton and David M. Godden, "Persuasion dialogue in online dispute resolution," vol., no., 2005. [Online]. Available: http://dx.doi.org/10.1007/s10506-006-9014-0

[54] Chris Reed and Glenn Rowe, "Reed & rowe, araucaria:argument diagramming and xml araucaria: Software for puzzles in argument diagramming and xml."

[55] Werner Kunz and Horst W. J. Rittel and We Messrs and H. Dehlinger and T. Mann and J. J. Protzen, "Issues as elements of information systems," Tech. Rep., 1970.

[56] Simon Buckingham Shum, "Cohere: Towards Web 2.0 Argumentation," in *COMMA*,

[57] Severin Isenmann and Wolf D. Reut, "IBIS - a Convincing Concept ... But a Lousy Instrument?" in *Symposium on Designing Interactive Systems*,

[58] Jonassen, David and Remidez, Herbert, "Mapping alternative discourse structures onto computer conferences," in *Proceedings of the Conference on Computer Support for Collaborative Learning: Foundations for a CSCL Community*, ser. CSCL '02. [Online]. Available: http://dl.acm.org/citation.cfm?id=1658616.1658650

[59] Timothy Chklovski, "User interfaces with semi-formal representations: a study of designing argumentation structures," in *In Under Review for the Intelligent User Interfaces Conference*,

[60] Lowrance, John D and Harrison, Ian W and Rodriguez, Andres C, "Capturing analytic thought," in *Proceedings of the 1st international conference on Knowledge capture.*

[61] Heng, M. S. H. and De Moor, A., "From habermas's communicative theory to practice on the internet," vol., 2003.

[62] Nikos Karacapilidis and Euripides Loukis and Stavros Dimopoulos, "A web-based system for supporting structured collaboration in the . . ." in *IN THE PUBLIC SECTOR. EGOV 2004*,

[63] Rahwan, Iyad, "Mass argumentation and the semantic web," vol., no., 2008.

[64] Park, Souneil and Kang, Seungwoo and Chung, Sangyoung and Song, Junehwa, "Newscube: Delivering multiple aspects of news to mitigate media bias," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ser. CHI '09. [Online]. Available: http://doi.acm.org/10.1145/1518701.1518772

[65] Travis Kriplean and Michael Toomim and Jonathan T. Morgan and Alan Borning and Andrew Ko, "Is this what you meant?: promoting listening on the web with reflect," in *CHI Conference on Human Factors in Computing Systems, CHI '12, Austin, TX, USA - May 05 - 10, 2012*, [Online]. Available: http://doi.acm.org/10.1145/2207676.2208621

[66] Faridani, Siamak and Bitton, Ephrat and Ryokai, Kimiko and Goldberg, Ken, "Opinion space: A scalable tool for browsing online comments," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ser. CHI '10. [Online]. Available: http://doi.acm.org/10.1145/1753326.1753502

[67] T. Kriplean and J. Morgan and D. Freelon and A. Borning and L. Bennett, "Supporting reflective public thought with ConsiderIt," in *CSCW*,

[68] Figallo, Cliff, *Hosting Web Communities - Building Relationships, Increasing Customer Loyalty, and Maintaining a Competitive Edge*. New York: Wiley Computer Publishing, 1998.

[69] ed. Boston, MA, USA: Addison-Wesley Longman Publishing Co., Inc., 2000.

[70] Preece, Jenny, *Online communities: Designing usability and supporting socialbilty*. John Wiley & Sons, Inc., 2000.

[71] Young, Margaret Levine and Levine, John, *Poor Richard's Building Online Communities: Create a Web Community for Your Business, Club, Assocation, or Family*. Top Floor Publishing, 2000.

[72] Ortega, Felipe and Gonzalez-Barahona, Jesus M. and Robles, Gregorio, "On the inequality of contributions to wikipedia," in *Proceedings of the Proceedings of the 41st Annual Hawaii International Conference on System Sciences*, ser. HICSS '08. [Online]. Available: http://dx.doi.org/10.1109/HICSS.2008.333

[73] Voss, Jakob, "Measuring wikipedia," 2005.

[74] Panciera, Katherine and Halfaker, Aaron and Terveen, Loren, "Wikipedians are born, not made: A study of power editors on wikipedia," in *Proceedings of the ACM 2009 International Conference on Supporting Group Work*, ser. GROUP '09. [Online]. Available: http://doi.acm.org/10.1145/1531674.1531682

[75] Bishop, Jonathan, "Increasing participation in online communities: A framework for human–computer interaction," vol., no., 2007.

[76] Brandtzæg, Petter Bae and Heim, Jan, "Explaining participation in online communities," 2009.

[77] Nonnecke, Blair and Preece, Jenny and Andrews, Dorine, "What lurkers and posters think of each other," in *Proceedings of the Proceedings of the 37th Annual Hawaii International Conference on System Sciences (HICSS'04) - Track 7 - Volume 7*, ser. HICSS '04. [Online]. Available: http://dl.acm.org/citation.cfm?id=962755.963094

[78] Ludford, Pamela J. and Cosley, Dan and Frankowski, Dan and Terveen, Loren, "Think different: Increasing online community participation using uniqueness and group dissimilarity," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ser. CHI '04. [Online]. Available: http://doi.acm.org/10.1145/985692.985772

[79] Fugelstad, Paul and Dwyer, Patrick and Filson Moses, Jennifer and Kim, John and Mannino, Cleila Anna and Terveen, Loren and Snyder, Mark, "What makes users rate (share, tag, edit...)?: Predicting patterns of participation in online communities," in

Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work, ser. CSCW '12. [Online]. Available: http://doi.acm.org/10.1145/2145204.2145349

[80] Yla R. Tausczik and James W. Pennebaker, "Participation in an Online Mathematics Community: Differentiating Motivations to Add," in *CSCW*,

[81] Huberman, Bernardo A and Romero, Daniel M and Wu, Fang, "Crowdsourcing, attention and productivity," *Journal of Information Science*, 2009.

[82] Wu, Fang and Wilkinson, Dennis M and Huberman, Bernardo A, "Feedback loops of attention in peer production," vol., 4.

[83] E. Gilbert, "Widerspread underprovision on Reddit," in *CSCW*,

[84] Cliff Lampe and Erik W. Johnston and Paul Resnick, "Follow the reader: filtering comments on slashdot," in *CHI*,

[85] Hogarth, R.M., *Judgement and Choice: The Psychology of Decision*, ser. A Wiley-Interscience publication. Wiley, 1987. [Online]. Available: http://books.google.de/books?id=dN6tB3Z552gC

[86] Stemler, Steven E, "A comparison of consensus, consistency, and measurement approaches to estimating interrater reliability," vol., no., 2004.

[87] Wen-Feng Hsiao and Hsin-Hui Lin and Te-Ming Chang, "Value-Based Consensus Measure on Verbal Opinions," in *HICSS*, 2005.

[88] Hiltz, Starr Roxanne and Johnson, Kenneth and Turoff, Murray, "Experiments in Group Decision Making Communication Process and Outcome in Face-to-Face Versus Computerized Conferences," vol., no., 1986.

[89] Siegel, Jane and Dubrovsky, Vitaly and Kiesler, Sara and McGuire, Timothy W, "Group processes in computer-mediated communication," vol., no., 1986.

[90] Walther, Joseph B, "Computer-mediated communication impersonal, interpersonal, and hyperpersonal interaction," vol., no., 1996.

[91] Sharon Murrel, "Computer communication system design affects group decision making," in *CHI*, 1983.

[92] Nunamaker, J. F. and Dennis, Alan R. and Valacich, Joseph S. and Vogel, Douglas and George, Joey F., "Electronic meeting systems," vol., no., Jul. 1991. [Online]. Available: http://doi.acm.org/10.1145/105783.105793

[93] Hilmer, Kelly M. and Dennis, Alan R., "Stimulating thinking: Cultivating better decisions with groupware through categorization," vol., no., Dec. 2000. [Online]. Available: http://dl.acm.org/citation.cfm?id=1289640.1289646

[94] Axtell, Robert and Axelrod, Robert and Epstein, Joshua M and Cohen, Michael D, "Aligning simulation models: A case study and results," vol., no., 1996.

[95] Moss, Scott, *Editorial Introduction: Messy Systems-The Target for Multi Agent Based Simulation.* Springer, 2001.

[96] Edmonds, Bruce and Chattoe, Edmund, "When simple measures fail: Characterising social networks using simulation," in *Social Network Analysis: Advances and Empirical Applications Forum. Oxford, UK*, 2005.

[97] Gilbert, Nigel, "Open problems in using agent-based models in industrial and labor dynamics," vol., no., 2004.

[98] Carley, Kathleen M and Wallace, William A, *Computational organization theory.* Springer, 2001.

[99] Kevin A. Gluck and Paul Bello and Jerome R. Busemeyer, "Introduction to the special issue," vol., no., 2008. [Online]. Available: http://dx.doi.org/10.1080/03640210802473582

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

[101] 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. [Online]. Available: http://dl.acm.org/citation.cfm?id=1027014.1027040

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

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

[104] 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. [Online]. Available: http://doi.acm.org/10.1145/1631144.1631149

[105] Archak, Nikolay and Sundararajan, Arun, "Optimal Design of Crowdsourcing Contests." p., 200. [Online]. Available: http://dblp.uni-trier.de/db/conf/icis/icis2009.html#ArchakS09

[106] 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. [Online]. Available: http://doi.acm.org/10.1145/1386790.1386829

[107] 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*, [Online]. Available: http://doi.acm.org/10.1145/1566374.1566392

[108] Smith, Robin, " , 2014.

[109] Hansen, Hans V., " Are there methods of informal logic?" in *Proceedings of OSSA 9, Argumentation: Cognition and Community,*, 2012.

[110] Neil Stuart Butt, "Argument Construction, Argument Evaluation, And Decision-Making: A Content Analysis Of Argumentation And Debate Textbooks," Ph.D. dissertation, Wayne State University, 2010.

[111] Groarke, Leo, " , 2013.

[112] David Hitchcock, "Informal logic and the concept of argument,"

[113] Walton, Douglas, *Argumentation schemes for presumptive reasoning.* Routledge, 2013.

[114] Woods, J. and Walton, D.N., *Fallacies: selected papers 1972-1982*, ser. Studies of argumentation in pragmatics and discourse analysis. Foris, 1989. [Online]. Available: http://books.google.de/books?id=dLIyAAAAIAAJ

[115] van Eemeren, F.H. and Grootendorst, R., *Argumentation, Communication, and Fallacies: A Pragma-dialectical Perspective.* L. Erlbaum, 1992. [Online]. Available: http://books.google.de/books?id=OFv5p3coL9sC

[116] Chris Reed and Douglas Walton, "Argumentation schemes in dialogue."

[117] Henry Prakken, "On the nature of argument schemes," 2010.

[118] Myerson, Roger B, *Game theory.* Harvard university press, 2013.

[119] Aumann, R.J., "Game Theory," Basingstoke: Palgrave Macmillan, 2008.

[120] Avinash Dixit and Barry Nalebuff, "Game theory," 2008. [Online]. Available: http://econlib.org/library/Enc/GameTheory.html

[121] T. L. Turocy and B. von Stengel, "Game theory," vol., Volume 2, pp 403–420, 2002.

[122] Osborne, M.J., *An Introduction to Game Theory.* Oxford University Press, 2009. [Online]. Available: http://books.google.de/books?id=_C8uRwAACAAJ

[123] Shoham, Yoav and Leyton-Brown, Kevin, *Multiagent Systems: Algorithmic, Game-Theoretic, and Logical Foundations.* New York, NY, USA: Cambridge University Press, 2008.

[124] Tanasijevic, Sanja, Böhm, Klemens, "A New Approach to Large-Scale Deliberation," in *The Third International Conference on Cloud and Green Computing (CGC 2013), Karlsruhe, Germany, Sept. 6– Oct. 10, 2013*, 2013.

[125] ——, "Towards Effective Structure-Based Assessment of Proposals and Arguments in Online Deliberation," *International Journal of Systems and Service-Oriented Engineering (IJSSOE).*

[126] Jurca, Radu and Faltings, Boi, "Minimum payments that reward honest reputation feedback," in *Proceedings of the 7th ACM Conference on Electronic Commerce*, ser. EC '06. [Online]. Available: http://doi.acm.org/10.1145/1134707.1134728

[127] Nolan Miller and Paul Resnick and Richard Zeckhauser, "Eliciting informative feedback: The peer-prediction method," in *Computing with Social Trust*, [Online]. Available: http://dx.doi.org/10.1007/978-1-84800-356-9_8

[128] Tanasijevic, Sanja, Böhm, Klemens, and Ehrhart, Karl-Martin, "Behavioral Strategies in Online Forums with Different Feedback Types," in *The 7th International Conference on Social Computing and Networking (SocialCom 2014)*, 2014.

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