Mediating debate through on-line large-scale argumentation: Evidence from the field

Ali Gürkan¹, Luca Iandoli, Mark Klein, Giuseppe Zollo

Article info

Article history:
Received 6 October 2008
Received in revised form 16 February 2010
Accepted 4 June 2010

Keywords:
Large scale argumentation
On-line deliberation
Collaborative technologies
Collective intelligence

Abstract

Web 2.0 technologies, such as forums and wikis, are enabling an explosion of global knowledge sharing through distributed large-scale conversations, but they seem to be less successful at supporting collaborative deliberation around complex and controversial questions. In order to cope with this limitation, many scholars have proposed to adopt on-line argumentation platforms to improve information visualization, organization and reuse. However, such research has mostly focused on the design of adequate argument-based knowledge formalisms. Less attention has been paid to the empirical analysis of actual interactions mediated by argumentation technology with reasonably large user communities. In this paper, we present an in-depth analysis of the data obtained in the empirical test of an argumentation platform where a 160-member community created, in 3 weeks, what is to our knowledge the largest single online argument map ever built (around 5000 posts). Our results show that (i) users were able to quickly and comprehensively explore and map the debate on the selected discussion topic; (ii) substantial moderation was needed to ensure that the argument map was well-organized and users were confident with the argumentation formalism; (iii) considerable out-of-the-map communication occurred, possibly as a way to allow for conversational flows inhibited by the argumentation formalism; (iv) formal rating of contributions favored exploration of the map, understanding the debate structure, and improving the quality of content.

© 2010 Elsevier Inc. All rights reserved.

1. Towards Internet-enabled collective intelligence

The Internet owes a great deal of its success to the emergence of collaborative technologies as forums, blogs, wikis, instant messaging, RSS, tagging, social network sites, and so on. Successful examples of virtual communities range from open source projects (e.g. Linux) to online encyclopedias (e.g. Wikipedia, with several millions user accounts). Their success has encouraged many companies to adopt collaborative technologies for business purposes and has inspired management and organizational science scholars to identify new models for organization design and governance [1,12,23,35,44,49,50].

On-line collaborative tools use cost-efficient technologies to harvest information from a large user community characterized by a considerable level of autonomy, openness, and rule-based self-governance. Some of these collaborative technologies, like forums and blogs, are used to mediate large, distributed conversations and even produce, in the case of wikis, collective knowledge outputs in the form of shared texts. While enabling effective information sharing and accumulation,
current tools get stuck when they are used for enabling discussions around complex systemic challenges involving high level of controversy, because of such notorious problems as “edit wars” and debate polarization. Such “systemic” challenges, including issues like climate change, energy, and security [26,36], are however the ones that most clearly and urgently need an effective means for incorporating the perspectives and expertise of many diverse individuals.

Researchers have developed on-line argumentation tools [2,3,9,21,28,34] to help to cope with these limitations. Such attempts generally assume that argumentation tools offer significant benefits in (1) improving knowledge access and reuse (by providing a logical rather than time-based organization for deliberations), (2) reducing the prevalence of the critical thinking pitfalls (by making argument chains more explicit and easier to check), and (3) avoiding important types of group deliberation failures (by removing the ability that users have, in time-based systems, to inappropriately dominate the attention allocation of the group). These research efforts have however focused mostly on the questions of formally representing and visualizing argument maps while less attention has been paid to the empirical analysis of how well argumentation technology fares when it is used to support reasonably large user communities engaged in controversial discussions.

We have implemented a large scale argumentation system, called the Deliberatorium, in order to explore this issue. This system combines ideas from argumentation theory with insights from currently available collaboration technologies to benefit from the advantages of argument mapping in contexts where large, diverse, and geographically distributed user communities want to deliberate about complex controversial problems. The system therefore integrates a set of core functionalities to support argumentation with more traditional web 2.0 technologies – such as discussion forums, instant messaging, e-mail alerts, and personal web pages – to support more unconstrained, out-of-the-map interaction and provide social visibility to users.

The paper is structured as follows. In the next section we discuss the limits faced by current technologies in supporting collective deliberation around complex problems. Then we outline the technical and organizational design of a large-scale argumentation platform that we believe can help to transcend these limitations. In the final part of this work we report empirical results and lessons learned from a field test of the Deliberatorium, in which 160 users created, in 3 weeks, what is to our knowledge the largest single online argument map ever built (around 5000 posts). The field results shed light on some key strengths and challenges of using argumentation tools for effectively mediating on-line debates.

2. Limitations of current Internet technologies to support collective deliberation

How well can current Internet technologies support collective deliberation? According to the widespread expectation and some empirical findings, by using Internet technologies, groups made up of sufficiently large number of capable members can leverage collective intelligence to outperform skilled but isolated individuals or small groups for such tasks as group problem solving [31], collective prediction [41] and deliberation on complex systemic problems [54]. There is, however, also a substantial literature about factors that can prevent groups from being smarter than their average member, factors that appear to be present in unstructured large-scale Internet-based conversations such as online forums [40]. The picture, as a result, is far from being simple.

Following de Moor and Aakhus [9] and Klein [25], we can classify current on-line deliberation technologies into three groups: sharing, funneling, and issue networking.

By far the most commonly used technologies, including wikis, blogs, and discussion forums, are what we can call sharing tools [21]. While such tools have been remarkably successful at enabling a global explosion of idea and knowledge sharing, they face serious shortcomings when applied for collective deliberation purposes. One is the signal-to-noise ratio. The content captured by forums is notorious for often being unsystematic and repetitive, making it hard for users to locate useful information. The content is also typically of highly variable quality, since online conversation tools do not inherently encourage or enforce any standard concerning what constitutes valid argumentation, so postings are often biased rather than evidence or logic-based. It becomes difficult, as a result, to separate the wheat from the chaff. Forum conversations do not scale well because it is virtually impossible for latecomers to make sense of and effectively join a conversation that has already started by other participants. While wikis are able to cope with some of the limitations above, they tend to fare poorly when applied to controversial topics, often leading to such phenomena as “edit wars”.

Funneling technologies include information aggregation markets (IAMs) and e-voting. When properly designed, e-voting systems can be successfully and easily employed for such tasks as assessing other users’ reputations, but tend to produce random or highly biased outcomes if they are used without market incentives or structured voting processes [21]. Good evidence for “wisdom of the crowd” [40,41] has however appeared in studies on on-line IAMs [55]. Participants in an IAM bet on the supposed best candidate and receive a financial incentive if their bet is correct. IAMs have been shown to constitute a reliable collective intelligence approach for such uses as predicting the winners of political elections (e.g. see the University of IOWA prediction market), but they cannot be used to deliberate collectively about complex problems for which the solution set is not pre-defined. IAMs also assume independence and lack of interaction among the participants, which is rarely satisfied in many contexts like organizations or work teams. When the independence requirement is not satisfied, prediction markets can produce biased outcomes [8]. The reason for the good performance of prediction markets relies first on the simplicity of the problems they can deal with (there is a known and limited set of possible alternatives), and second on the presence of market incentives that motivate individuals to search for more information and prefer rational choices. This appears to represent a major difference from other on-line collaborative communities, like Wikipedia or the open-source movement, in which many different kinds of both extrinsic (e.g. payments) and mainly intrinsic incentives (reputation, reciprocity,
entertainment, voluntary contributions, etc.) are at work. While funneling technologies can be effective at aggregating information provided by large groups made up by diverse and independent individuals, they do not help groups to develop a shared understanding of a problem, and provide no visibility on the process through which members retrieve and create knowledge. Thus, funneling tools appear to be useful for harvesting and aggregating information that a group already possesses rather than for the creation of new knowledge through deeper understanding, higher awareness or for reframing problems and solutions.

Issue networking tools, also known as argumentation or rationale capture technologies, try to fill this gap by helping groups collaboratively define networks of issues (questions to be answered), options (alternative answers for a question), and arguments (statements that support or detract from some argument or idea), typically in the form of visual maps (see Fig. 1 for an example). Such tools can help to make deliberations, even complex ones, more systematic and complete. The central role of argument entities is, in addition, expected to encourage critical thinking, by implicitly requiring that users express the evidence and logic in favor of the options they prefer. The results are captured in a compact form that helps users to understand and reflect on what has been discussed to date and, if desired, add to it without needless duplication, enabling increased synergy across group members as well as cumulativeness across time.

Current issue networking systems do face a critical shortcoming however. They have been applied to support only small-scale deliberations. Most issue networking systems are used to support physically co-located team meetings where a single facilitator captures free-form discussions in the form of a commonly viewable argumentation map. A few have been used to enable non-facilitated deliberations with physically distributed participants over the Internet. In both cases the number of authors working together on any given map has been small, on the order of 10 or less, far less than what is implied by the vision of large-scale collective intelligence presented in this paper.2

There has been, consequently, almost no empirical analysis of the actual use of argumentation systems by reasonably large communities. Moreover, the developers of issue networking tools have focused their efforts mostly on the design of adequate knowledge formats, while not much research has been carried out to observe participants’ reaction and the actual usage of such formats (with a few exceptions in the field of e-learning or for off-line tools). This paper aims at making a contribution in filling that gap.

2 The only exception we know of (the Open Meeting Project’s mediation of the 1994 National Policy Review) supported hundreds of contributors but was effectively a comment collection system rather than a deliberation system, since the participants were predominantly engaged in offering reactions to a large set of pre-existing policy documents, rather than interacting with each other to create new arguments or options.
In the next section, we briefly outline our approach to the design of a large scale on-line argumentation collaborative platform, while in Section 4 we report the empirical findings obtained through a field test of the tool.

3. The design of a large scale argumentation platform

Designing a large-scale argumentation platform requires specifying both how the community collectively generates content (i.e. the organizational design) and the formalism by which contributions can be expressed. We address these issues in the sections below.

3.1. Organizational design

The design of a virtual community involves dealing with several governance issues such as the design of incentives to attract and motivate a critical mass of users, and the organization of collaborative processes.

The identification of a clear mission is crucial for attracting the right kind of committed users because deliberations focus on specific topics. In general, attracting a large number of committed and diverse users is desirable. The diversity and ability of users however, have to be equally balanced [31]. Attracting diverse expertise for collaborative purposes may not be an easy task, since experts may decide to retain their knowledge or may require adequate incentives to participate such as the awards offered for best solutions on the web portal Innocentive [33]. Benkler [1] identifies three classes of incentives: monetary, hedonic and socio-psychological rewards, but he shows that the choice of the right mix of incentives may depend on many variables including the nature of tasks.

Since we ran our test with participants who in some cases knew each other, several countermeasures and incentives were set up to limit the potentially negative effects of limited scale and participation as well as the presence of social and information pressure usually absent or limited in large Internet communities. In particular, we used several extrinsic incentives such as minor awards and five scholarships for the best participants thanks to the support of a sponsor, with the aim of attracting a sufficient number of users and improving the quality of contributions. This choice however made our experimental context different from many real on-line collaborative venues like for instance open source communities, in which intrinsic incentives are prevalent [38].

In terms of labor division and coordination, collaborative argumentation takes place in the platform through a simple process in which issues and ideas submitted by users are supported and attacked by arguments from other users through a process whose aim is to uncover the chains of pros and cons behind each idea. This knowledge is structured and organized through argument mapping and visualization.

A critical issue is the availability of adequate argument mapping skills. Since participation is open and voluntary, one cannot take for granted that a sufficient number of skilled argument mappers will be available, nor that common users will spend enough time and effort to learn to use the formalism properly. The solution we adopted for this study was to introduce a moderator corps. Moderators help to ensure that posts follow the basic rules of structured argumentation and are properly located. Their role is that of a honest broker, making sure the material is structured in a way that maximizes its usefulness for users of the system. Moderators are also expected to provide feedback to users to help them learn how to use the tool correctly. They can be selected from among existing users on a meritocratic basis, through a bottom-up self selection process based on reputations assigned by the community.

Consequently, in the Deliberatorium virtual community, there were three types of user: moderator, author and reader/voter. Moderators were in charge of filtering out noise and rejecting off-topic posts. They were also responsible for ensuring that the argument map was well-structured, i.e. all posts were properly divided into individual and non-redundant issues, ideas, and arguments, and were located in the relevant branch of the argument map. This involved classifying and sometimes editing posts, offering suggestions to authors, aggregating similar arguments, and occasionally re-organizing the overall argument map so that related topics are grouped into the same branch. A team of four student moderators was selected and trained in argument mapping before the test. One of the authors assumed the moderator role as well. The top down choice was made because the limited time window for the experiment would not have allowed us to observe the emergence of a self-organized hierarchy. The on-line argumentation process developed as follows:

1. Authors posted and edited issues, ideas, and pro/con arguments and produced an argument map similar to that in Fig. 1. While issues and ideas could be posted only as single short sentences, arguments were posted using an online form that helped users to structure their post in the form of an argument (constituted by a conclusion, an argument scheme and grounds) and also allowed them to add web links and documents. The form was designed to help users to unbundle their contributions into issues, ideas and arguments according to the proposed argument formalism (see Fig. 2). We established a single authorship rule: nobody, except moderators, could edit a post authored by someone else.
2. All users (including moderators, authors and readers) could rate arguments and ideas. As in wiki talk pages, they could also send comments to authors through threaded discussion forums associated with each post. Rating was anonymous.
3. Posts were initially given a status of “pending”, and could only be certified by moderators. Until a post was certified, it could not be rated and nobody, except its author, could link any other post to it. The rule was that only certified posts would have appeared in the final, publicly available version of the argument map and it was not possible to attach
additional posts to uncertified posts. This helped to limit useless proliferation that can be observed in discussion forums in which bad or deliberately provocative posts can trigger long but usually low value-added discussion threads. Moderators also left comments, edited, moved, trashed and classified posts. Usually moderators would leave a comment to explain changes. Authors would receive an alert email when their post was modified or trashed (but the trash was never emptied, so posts could always be recovered if necessary).

3.2. Formalism

In the last decade, several researchers have invested considerable effort in developing human-centric argumentation tools [2,7,9,21,28,34,42] with the aim of enabling “mass conversations” that lead to the emergence of more plausible, convincing and shared conclusions about a given topic while allowing the coexistence of conflicting perspectives. Such approaches depart from the more formal approaches taken by artificial intelligence researchers [10] by utilizing argument representations that are less formally expressive but also substantially easier for people to understand and use.

Our approach integrates three previously-developed human-centric argumentation formalisms: IBIS (Issue Based Information System, [7]), the Toulmin framework [45] and the concept of argument schemes proposed by Walton [52,53] recently adopted by others [3,34]. The proposed representation aims at being as close as possible to everyday argumentation, keeping formalization at a minimum without losing the structuring power of arguments.

The IBIS approach represents arguments by using three basic elements: Questions (called ‘Issues’ in the Deliberatorium) which pose a problem to be solved, Ideas which offer possible solutions, and pro/con Arguments which support or reject an idea or another argument. In the IBIS framework, arguments develop as trees as in Fig. 1. Several tools for argument mapping have been developed, such as Belvedere [42] and Compendium [39], but applications have largely been limited to supporting individuals or small scale, physically co-located groups, the latter usually requiring the presence of a trained facilitator. An interesting aspect of the IBIS formalism is that it fosters the development of argument maps by supporting a conversational interaction thanks to the availability of question and idea nodes. Consequently, a good IBIS map is extremely sensitive to the way questions are chosen and formulated [7].

While in IBIS pros and cons are atomic components, the Toulmin argumentation framework proposes a more fine-grained representation of arguments based on the identification of functional roles played by different argument components. According to Toulmin, an argument is a sequence of interconnected affirmations (claims) that establish the content and the strength of the position of the orator [15]. Claims can be classified into different functional categories, namely the key claim, or conclusion of an argumentation, the grounds, such as the facts and opinions offered to support a key claim; and the warrants, i.e. justifications for inferring the claim from the provided grounds.

The role of warrant is crucial in effective argumentation. Helping users uncover implicit warrants may help them to identify flawed arguments or discuss the taken-for-granted assumptions contained in hidden warrants. On the other hand, there is evidence that warrants tend to be hard to recognize and elicit in both written and conversational arguments [19,51,29]..
and the concept is hard to grasp for inexperienced users [30]. A potentially easier alternative for at least partially capturing arguments is the concept of the argument scheme proposed by Walton [53]. Schemes represent stereotypical, commonly-used ways of drawing inferences [34] that can be considered acceptable in the absence of complete information. Though several structures and taxonomies of schemes have been proposed [32], Walton’s exposition is very appealing since his classification is drawn from the everyday use of arguments. A second desirable characteristic of Walton’s schemes is that a set of critical questions is assigned to each scheme, enabling contenders to identify the weaknesses of an argument and potentially attack it (see some examples in Table 1).

Walton’s schemes can be used (i) by readers, to recognize and classify arguments proposed by other users and to check if the critical questions are adequately answered and (ii) by authors, to check if their arguments are defendable with respect to the critical questions.

By merging the IBIS, Toulmin and Walton formalisms, our aim is to propose a representation that helps people to distinguish the inputs to an argument (i.e. facts, evidence, shared opinions, values, etc.) and the reasoning scheme through which conclusions are obtained from these inputs. This critical distinction is made for two reasons: (i) to encourage evidence-based reasoning; and (ii) to induce users to consider the validity of an argument by assessing the credibility of both the grounds and the reasoning scheme. In other words, an argument’s pitfalls can be found in the supporting evidence, in the scheme, or both.

### 4. Empirical test of the platform

#### 4.1. Test objectives

A first test of the Deliberatorium was performed in December 2007 at the University of Naples Federico II (Italy) with a community of about 160 graduate students, which was asked to deliberate on the topic given: “the future of biofuels”.

The Deliberatorium is a Common Lisp application developed on top of cl-http, an open source web server developed at MIT (http://www.cl-http.org:8001/). It provides a simple and consistent web-based user interface that allows users to navigate and edit the argument map as well as communicate with each other. The system’s capabilities are made accessible via a set of tool icons arrayed across the toolbar at the top of the page (Fig. 3).

The key capabilities of the Deliberatorium represent a combination and extension of the key features found in sharing, funneling, and issue networking technologies. Such an integration is a distinctive feature of the Deliberatorium and aims at facilitating large scale interaction (see [17,25] for a more detailed discussion of the design of the Deliberatorium).

The primary aim of the experiment was to observe users’ reaction to the platform and especially to the shift from the conversational, time-based communicative style, common in such existing collaboration tools as forums, to the logic-based approach required by argumentation. Second, we wanted to evaluate the performance of Deliberatorium in terms of knowledge accumulation and organization as well as user participation and satisfaction. Finally, we intended to use the test results to identify improvements and revisions in the design of the tool.

### Table 1
Examples of argument schemes. (our adaptation from Walton).

<table>
<thead>
<tr>
<th>Argument scheme</th>
<th>Argument structure</th>
<th>Critical questions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Expert opinion</td>
<td>Ground: E is an expert in the domain A is in</td>
<td>How credible is E (reliable, free of conflict of interests, authoritative, etc.)? Is E an expert in the field A is in? Is E’s assertion based on evidence?</td>
</tr>
<tr>
<td>Popular opinion</td>
<td>G: A is generally accepted as true W: Believe what is generally accepted as true</td>
<td>What evidence (e.g. polls) supports that A is generally accepted? Even if A is generally accepted, are there any reason for doubting it is true?</td>
</tr>
<tr>
<td>Analogy</td>
<td>G: Case C1 is similar to case C2, A is true in C1 W: repeat things that have proven to work well in the past</td>
<td>Are there differences between C1 and C2? Was A correct (true) in C1? Is there any other case C3 similar to C1 in which A was not correct/true?</td>
</tr>
<tr>
<td>Causal (contains variant the argument form consequences and the slippery slope argument)</td>
<td>G: there is a positive correlation between A and B W: find out causal relationships between things happening together</td>
<td>Is the correlation supported by credible evidence? Is the correlations due to coincidence? Could there be some factor C causing both A and B? Are there any other consequences to A that should be taken in the account? What evidence support that given A, B will really occur? What factors can prevent the causal chain to happen and how much are they probable? What is the weakest link of this chain How much is the probability that the chain will actually start?</td>
</tr>
</tbody>
</table>
In the following we report the outcomes of some statistical analysis of the data contained in the Deliberatorium database. The database did not only host the content posted by users but also tracked virtually every user action (all told, about 110,000 time-stamped events). Given the aims and the novelty of the experiment as well as the lack of existing theories about on-line large-scale argumentation, our results are mainly descriptive and do not aim at testing specific hypotheses. In general terms, our objective is to observe, in the field, whether argumentation can effectively mediate on line discussions with large groups in a sustainable way and to assess users’ reaction to the constraints on the content format and interaction process imposed by argumentation.

4.2. Experimental setup

The students involved in the experiment were all part of a same large class from a graduate program in Industrial Engineering, age 23–25, 55% male. Students selected from that class helped to coordinate and manage the experiment, and were required as a result to deal with social pressures from their fellow students. All these circumstances made the context different from a fully open online community and represent a significant limitation of this study. On the other hand, going large scale in these early steps within an uncontrolled experimental setting might not have attracted a critical mass of users and would have prevented us from having a direct contact with them, which proved to be very useful for debugging, improving and upgrading the software based on user feedback.

The evaluation proceeded in three phases:

Phase 1: Preparatory work.
Phase 2: A three weeks period, in which students were requested to populate the Deliberatorium argument map.
Phase 3: Data analysis.

In the preparatory phase, the students had four 2 h seminars from external experts about:

2. Argumentation, with focus on the IBIS approach.
3. Major issues in energy governance with a country focus on Italy and UE policies.

The students were also given a few reading materials: two newspaper and magazine articles about the topic and the IBIS manual available at http://touchstone.com/wp/IBIS.html [6]. As a discussion topic we chose “the future of biofuels”. The
criteria we used to select the topic were: (1) it had to involve a current debate about a complex systemic challenge (such as global warming), (2) it had to be focused enough to help students not to get lost into a too wide a debate, considering they had limited time, attention and expertise, and (3) it had to be controversial and multifaceted so that the community could explore possible different solutions and conflicting perspectives.

4.3. Data analysis

4.3.1. Database preparation and setting

The Database assigned a user ID for every user name. Some users had had problems when they first started using the platform. Therefore, some created multiple accounts. Totally 184 user names are identified in the database 12 of which were assigned to the moderators, the administrator and accounts created for visitors. Even though some users had multiple user accounts, they mostly preferred the same user name. This created multiple user ID's with a unique user name which helped us to distinguish duplications. To increase the reliability of data, student names taken from the class registration list were compared with user names used in the platform. After several controls 11 user names were deleted because they represented duplicate accounts, and two others were deleted since they did not represent any activity or they had an unusual pattern. Overall, the number of accounts considered was 159.

4.3.2. Users' participation

We observed a very high level of user participation as we achieved thousands of posts in just a few days. Remarkably, the Deliberatorium was active almost 24 h per day, except for a hiatus between roughly 3 and 6 am. In two weeks, the users posted nearly 3000 issues, ideas and arguments (of which roughly 1900 were eventually certified) in addition to over 2000 comments (see Table 2).

Though students' participation may have been influenced by their perception that the experiment was a course task for which they could be evaluated by their professor, their informal face-to-face and on-line comments posted on the Deliberatorium as well as on a threaded discussion forum run independently by a student's association web site, showed that they found the experiment interesting and appreciated the innovative characteristics of the tool.

4.3.3. Breadth of coverage

The breadth of coverage as well as the efficiency of the platform in terms of knowledge accumulation was quite good: a non-expert community of students was able to create a remarkably comprehensive map of the current debate on biofuels in just a couple of weeks, exploring topics ranging from technological issues to environmental, economic and sociopolitical impacts of the widespread diffusion of biofuels. Moreover, the proportion of out-of-topic posts was negligible – about 0.1%.

In order to have a more accurate idea of the subtopics addressed by the users as well as the overall breadth of coverage of the debate, we performed a cluster analysis on a sample of 690 randomly selected arguments. We generated a first list of keywords by using a tag cloud generator available on http://www.wordle.net starting from the text of the 690 posts contained in our database. Such list was then revised in order to eliminate irrelevant words and group slightly different wordings of a term under a same label. The refined keywords list contained 162 terms. Then we assigned to each argument a vector of 162 dimensions each measuring the frequency with which a given term occurred in the argument itself. Finally we applied Ward's hierarchical clustering algorithm to identify a number of relevant topic clusters. After a few iterations we identified and eliminated from the sample some outliers that were responsible for introducing distortions in the definition of the clusters. Then we used agglomeration coefficients to measure the change in heterogeneity that occurs when the number of potential clusters is reduced from $n + 1$ to $n$. With the elimination of outliers resulting from the clustering process, the number of posts taken into consideration decreased to 661. Eventually the solution with $n = 6$ proved to be the one which ensured the best compromise in terms of heterogeneity measures and coherent semantic discrimination. Though cluster analysis involves a certain level of subjectivity in the interpretation of the results, it gives a more accurate idea about the content developed by users compared to a subjective qualitative analysis. In our case we obtained the following clusters:

- (1) Technologies for the production of biofuels (65.5% arguments).
- (2) State of the country (Italy) in the production of biofuel (8%).
- (3) Impact of biofuels on agriculture and food production (9%).

Table 2

<table>
<thead>
<tr>
<th>Post type</th>
<th>Number of posts</th>
<th>Number of certified posts</th>
<th>% Certified</th>
</tr>
</thead>
<tbody>
<tr>
<td>Issue</td>
<td>242</td>
<td>89</td>
<td>5</td>
</tr>
<tr>
<td>Idea</td>
<td>962</td>
<td>452</td>
<td>24</td>
</tr>
<tr>
<td>Pro</td>
<td>1488</td>
<td>1045</td>
<td>55</td>
</tr>
<tr>
<td>Con</td>
<td>402</td>
<td>325</td>
<td>17</td>
</tr>
<tr>
<td>Comment</td>
<td>2009</td>
<td>n/a</td>
<td>n/a</td>
</tr>
<tr>
<td>Total</td>
<td>5003</td>
<td>1911</td>
<td>100</td>
</tr>
</tbody>
</table>
(4) Impact on greenhouse gases emission (12%).
(5) Biofuels and VS hydrogen (1.5%).
(6) Impact and role of biofuels on global energy policies (4%).

It is important to note that the above classification is crisp, so it does not consider multiple memberships, though a given argument might belong with different degrees to more than one cluster. The list of clusters shows that users touched relevant and topical subjects in the current debate on biofuel. The focus on technological issues is possibly due to the fact that users were engineering students.

4.3.4. Argumentation use
As expected, at the beginning of the experiment most users really did not grasp the argumentation logic. Many users adopted a kind of forum conversational style instead, in which they tended to publish posts as news articles (e.g. “France creates incentives for biofuel”) rather than as argument entities. Common mistakes were: difficulty in distinguishing between ideas and arguments, the tendency to put multiple arguments into a single argument post, linking arguments to logically irrelevant locations in the argument map, issues and ideas proliferating without any associated pro/con arguments, and difficulties in selecting the right kind of scheme for arguments. After a while we observed an improvement in the use of the platform, as users developed confidence, profited from moderator feedback, and learned to use the tool.

The level of direct debate was moderate. Users did attach many arguments to each other’s posts (62% of all posts were arguments and 67% of these arguments were attached to posts authored by someone else) but the great majority of all arguments (78.3%) were pros rather than cons. The depth of the argument trees was however relatively small. Most arguments (85%) were attached directly to ideas, with the remainder attached to other arguments. This relative dearth of debate may have been an outcome of students’ reluctance to criticize the contributions of their peers. Other possible explanations include: (i) inertia deriving from the predominant use of forums and wikis, (ii) the short time window available to learn the tool, (iii) the lack of specific expertise of the students on the topic, leading to fast content saturation and inability to explore specific subtopics in-depth; (iv) the use of individual awards and the single authorship rule which may have fostered competition over collaboration. These factors can explain a certain level of redundancy, an emphasis on authoring rather than debating to maximize individual exposure, as well as the fear of being involved in sub-discussions potentially characterized by a high level of conflict.

The Spearman correlations among different post types provide two interpretations (see Table 3). Comments are not very related to any post type while ideas and pros are highly correlated. The former might be due to the looser structure of comments which can be attached to any kind of post. They can represent an addition to the forum or a citation for a post. For better statistics, comments can be distinguished in separate subcategories in future experiments. Finally, the high association between pros and ideas is not surprising given that in a sample of students where people probably knew each other, users posted pros much more than cons. Since the depth of the deliberation was not very large (97% of posts are located in the first four levels of the tree), the discussions grew more breadth-wise than depth-wise, thereby creating new issues and ideas which are traditionally followed by pros and cons. The relatively low value of the correlation between con and pro or con and idea may be an indicator of reluctance to argue against others’ posts, probably due to social pressure.

4.3.5. Moderators’ role and behavior
Other important implications concern moderators, more precisely the community governance. Moderators played a crucial role by leading the community. They supported users with comments and suggestions and by ensuring a logically-organized argument map; they helped users to rapidly locate the contexts where their piece of knowledge can best be linked to. For these reasons it is crucial to have enough moderators to ensure fast certification and timely reorganization of the argument map. With the existing data we can roughly estimate the requisite number of moderators. A cadre of from 2 to 5 moderators (the number varied from day to day according to their other commitments during the course of the experiment) was able to more or less keep up with 160 authors, but only by dint of an unsustainably heavy investment of their time. We estimate that a more realistic time commitment would require between 5% and 10% of the active users to be moderators.

<table>
<thead>
<tr>
<th>Table 3</th>
<th>Correlations between post types (significance levels provided in parentheses).</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Pro</td>
</tr>
<tr>
<td>Pro</td>
<td>1.00</td>
</tr>
<tr>
<td>Con</td>
<td>1.00</td>
</tr>
<tr>
<td>Issue</td>
<td>1.00</td>
</tr>
<tr>
<td>Idea</td>
<td>1.00</td>
</tr>
<tr>
<td>Comment</td>
<td></td>
</tr>
</tbody>
</table>
4.3.6. Analysis of user activity

The objective of the statistical analysis was to exploit the data recorded into the Deliberatorium database to analyze user transactions. We characterized transactions in terms of user activities (actions), outputs (posts) and an individual performance indicator (efficiency) reported in Table 4. ‘Issue’, ‘Pro’, ‘Con’, ‘Idea’ and ‘Comment’ correspond to the number of posts authored by a user with respect to the post type. ‘All Posts’ is the sum of all these post types.

As far as activities, we included elementary actions like “view”, “rate”, and “create”. The event records include information on user views of posts. This allowed us to separate views of posts authored by the user (‘Viewed Own’) from those authored by authors other than the user (‘Viewed Else’) while these two variables together constituted the ‘Total View’ variable. When a user authors a post, he/she attaches it to a related other, which is also registered. Therefore we can see how many posts were added to posts authored by the user him/herself (‘Incoming Posts’) or to one authored by someone else (‘Outgoing Posts’). ‘Net Outgoing Posts’ is the difference of these two variables for each user, which stands for to what extent a user is oriented towards contributing in community’s or to his/her own arguments. We also measured how many times users entered ratings for posts. ‘Times Rated Others’ stands for the number of votes given to other users. The two nominal variables used classify users in terms of their rating patterns. ‘Rater 4 Category’ includes four user types: those who do not vote, do vote but only for him/herself, vote only for others and both. The ‘Binary Rater Category’ simply separates those who voted from those who did not.

Efficiency variable gives an idea of a user's performance. It is computed as the number of contributions made by a user, divided by the total number of views they make (extent of use). The contributions include posts authored and ratings. Ratings are essential for the functioning of the community, so they are taken into account as well.

Initial analysis revealed that none of the numerical variables was normally distributed (Kolmogorov–Smirnov and Shapiro–Wilk statistics for variables are respectively as follows: Issue: .27, .72; Idea: .14, .80; Pro: .14, .92; Con: .21, .76; Comment: .28, .62; All Posts: .10, .93; Viewed Own: .23, .64; Viewed Else: .17, .82; Total View: .17, .77; Incoming Posts: .15, .87; Outgoing Posts: .11, .92; Efficiency: .15, .86; Net Outgoing Posts: .12, .92; Times Rated Others: .30, .53; all with df of 159 at significance at .00 level which stands for the violation of the normality assumption).

As shown by the histograms in Fig. 4, we can see that the distribution of variables resemble the power law distribution which has been found to be typical of most on-line communities [27,14], but show sometimes a thick middle.

To analyze this aspect a Pareto analysis of event types has been done. Each user is associated with how many times s/he conducted the three major events, namely “create” (created a post), “view” (viewed a post, an author’s account, etc.) and “rate” (voted for a post). For each event type, we identified “top users”, i.e. people who are among the top 80% most frequent performers of an event. The largest group is formed by 62 users out of 159 (39% of the sample) who are not involved in any one of the top user groups; 31 students (19%) are present at one of the top create, view or rate user groups. This figure decreases to 27 students (17%) for users active at two event types. Finally 39 students, making up a quarter of the sample, are top users in all three event types. So there is a core group fueling the deliberation while a larger part does not get involved in activities to a remarkable extent.

For the “create” event, the customary 80% limit is hit when the top 84 users are accumulated (54% of the sample). The same limit is reached at 74 (48%) and 44 (29%) students for the “view” and “rate” events respectively. As these figures show, the real difference results from the rating activity, implying that raters might be the key to understand participation governance. The results of the Pareto analysis induced us to go more in depth on the analysis of user behavior with respect to rating. Rating can be a crucial activity in deliberation since it presumes analysis of content in the map. In real online communities rating is usually a very popular tool and is used as an input to more or less sophisticated and reliable trust and reputation systems, like those used in e-commerce websites to assess buyers’ and sellers’ trustworthiness (see [20] for a review).

---

3 With the word 'post' we refer to all posts except comments, i.e. issue, idea, pro and cons. Comments are neglected because it is the only post type which does not have rules defining the structure to follow while authoring.
Some statistical tests about ratings and rating behavior are conducted as presented below. Because none of the variables enlisted above are normally distributed, non-parametric tests are applied for further examinations.

4.3.7. Rating activity

Can rating activity be used as a proxy for the intensity and quality of contribution? To answer this question, raters are compared to non-raters in terms of the number of entries they authored to see whether they exhibit discrepancies in terms of participation. A preliminary visual analysis points out that the major difference is very likely to be between raters and non-raters (i.e. groups 0 vs. 1, 2, 3 defined for ‘Rater 4 Category’ variable in Fig. 5) because the group 1 (rated only his/her own posts) is formed by only five people while those who rated only for others (group 2) and rated both for others and themselves (group 3) do not seem to differ.

This is supported by the box-plot below where the total number of posts seems to differ for raters and non-raters, which correspond to 28 and 131 authors respectively (see Fig. 6). To verify this observation, the Mann–Whitney Test was done using the ‘Binary Rater Category’. This test shows that number of posts, number of views as well as the ‘Net Outgoing Posts’ differ significantly (see Table 5 below). The higher mean rank of raters in all three tests implies that they are superior to non-raters on these measures. The effect size\(^4\) is medium for ‘Net Outgoing Posts’; and large for ‘Total View’ and ‘All Posts’\(^5\). In the light of this test we can affirm that:

\(^4\) Effect sizes are computed according to the formula \(r = z/\sqrt{N}\) [11].
Raters use the platform more, and their contribution in populating it is higher than non-raters. Being a rater does not necessarily imply criticizing other authors without making a significant contribution. Even though the effect size is not very large, the difference between posts added to others’ posts versus the author’s own posts (see variable ‘Net Outgoing Posts’) is higher for raters. So raters not only evaluate other authors by their ratings, they also expand the deliberation in a constructive way, by adding posts to other authors’ posts.

The association between viewing (total view) and rating can be misleading. Because a rater gives his/her vote by visiting the post to be rated and this is registered as a view event in the database. The same holds also for posts created. Therefore a new variable, “surfer”, is created to compare raters against non raters with respect to the extent to which they navigated the map for the sake of exploring the debate. It is defined as:
Surfer = views of posts (total view) – number of posts authored (all posts) – number of rates.

This represents the extent to which an author uses and surfs the deliberation platform without rating. The distribution is again non-normal with $p: .00$ (df: 159) shown by the Kolmogorov–Smirnov and Shapiro–Wilk tests where z values are .18 and .75 respectively. Therefore Mann–Whitney test is used which yields a significant relationship with a large effect size ($r: .49$) (see Table 6). Given that raters have a higher mean rank than non-raters in surfing the platform; we can consider raters as more frequent users of Deliberatorium.

We also carried out a statistical test to check if there is a difference between regular and top raters. This question is partially addressed in the paragraphs above. As mentioned above, raters’ view measures might be artificially higher, because every rating is accompanied with viewing that item. Then a reasonable test can be to see whether the most active raters of other authors’ posts view their own posts. This indicates whether they just criticize others without considering the quality of their own arguments. So unlike the test presented above, this test considers only views of authors’ own posts. A positive result would mean that raters of other authors are more careful about their own contributions as well. However looking simply for an association between authors’ viewing their own posts and the times they vote for others can be misleading because viewing a post can simply be done upon creating it. Therefore to test the relation between author’s views of their own posts and the extent to which they vote for others, we look at the association between a new variable named self-check and the times an author voted for others where

\[
\text{Self Check} = \text{views of authors’ own posts (viewed own)} – \text{number of posts created times an author rated/her/himself.}
\]

The variable was not normally distributed ($p: .00$ both with Kolmogorov–Smirnov and Shapiro–Wilk tests; df: 159; z: .23 and .61 respectively). Therefore spearman correlation is used for observing the relation between ‘self check’ and the number of times a user rated others. It yields a high level of association where the correlation coefficient was .60 and significant at a .00 level.

Thus we can affirm that users, who rated others’ posts, view their own posts more frequently as well. This result implies that users who rate others are more prone to comparing their post with what others have said, i.e. they are perhaps more frequent users of Deliberatorium.

Table 5
Comparison of raters (N: 131) and non-raters (N: 28) in terms of ‘All Posts’, ‘Total View’ and ‘Net Outgoing Posts’ using Mann–Whitney Test results.

<table>
<thead>
<tr>
<th></th>
<th>All Posts</th>
<th>Net Outgoing Posts</th>
<th>Total Views</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Non-rater mean rank</td>
<td>Rater mean rank</td>
<td>Mann–Whitney U</td>
</tr>
<tr>
<td>Non-rater mean rank</td>
<td>34.54</td>
<td>89.72</td>
<td>561.00</td>
</tr>
<tr>
<td>Rater mean rank</td>
<td>57.13</td>
<td>84.89</td>
<td>1193.50</td>
</tr>
<tr>
<td>Mann–Whitney U</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Significance</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Effect size r</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total Views</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Non-rater mean rank</td>
<td>28.52</td>
<td>38.52</td>
<td>392.50</td>
</tr>
<tr>
<td>Rater mean rank</td>
<td>91.00</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mann–Whitney U</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Significance</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Effect size r</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 6
Comparison of raters and non-raters in surfing the map.

<table>
<thead>
<tr>
<th></th>
<th>Non-rater mean rank</th>
<th>Rater mean rank</th>
<th>Mann–Whitney U</th>
<th>Significance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Non-rater mean rank</td>
<td>30.96</td>
<td>90.48</td>
<td>392.50</td>
<td>.00</td>
</tr>
<tr>
<td>Rater mean rank</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mann–Whitney U</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Significance</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Effect size r</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
inclined to critical thinking (information search and verification). Since the overall number of raters is not small, this also represents an encouraging result: users were able to explore the debate following the map structure and compare their knowledge with that provided by others.

Finally we tested for a relationship among rating and efficiency. When rating behavior is compared against the efficiency of users, raters turn out to be more efficient. The mean rank obtained from the Mann–Whitney test increases for the active raters and the difference obtained is significant (p: .00) even though the effect size is low (r: .24) (see Table 7). So raters' efficiency is found to be higher than non-raters, showing that their higher contribution level is not the only factor that renders them valuable. They perform considerably better as well.

5. Discussion

In the following we discuss the limitations of the current work. We outline design implications as well as the lessons learned from this study.

5.1. Limitations of the study

The aim of the study was to observe, in the field, a medium-sized group of subjects using an argumentation platform in their deliberations. The argumentation platform aimed at mediating collective deliberations on controversial topics. The observations let us draw implications on how to improve the technical and organizational issues.

The study was exploratory in nature, mainly oriented toward finding design implications rather than testing specific hypotheses. Many more lessons almost certainly remain to be gleaned from the test data set. The Deliberatorium software recorded essentially every user interaction including every view, modification or rating of any post, so we have a complete time stamped record of the evolution of the argument map and what the users did while creating it, a database of over 110,000 distinct events. The results of a thorough analysis of this data will be presented in future publications.

The experiment involved a relatively small number of users by Internet standards, and the way students approached the experiment was distorted, no doubt, by the fact that they were co-located peers in another setting, i.e. in their course. They may also have believed that their grades would have been affected by their level of participation. Social pressure, coupled with the students' low expertise in the topic, may for instance have had a role in creating a shallow tree structure, limiting the number of cons compared to pros and reducing the number of poor ratings. The experiment also ran, perforce, over a limited time window.

Further evaluations will aim at removing these artificial constraints by assessing the platform with much larger, more open, and more intrinsically motivated user communities. Increased scale will require changes in design choices and user incentives as discussed in detail in Section 5.2. To show that the approach can scale, the study can be repeated by involving groups of students from different universities. In this way users of the community would be distributed geographically and have different cultures. We are also identifying other possible contexts for assessing and applying the Deliberatorium, ranging from problem solving within companies and professional communities of practice, to learning and education with communities of young students.

As discussed in Section 5.2, one major limitation was the disproportionate use of extrinsic incentives (awards for best participants, grading, etc.). It is highly probable that this, in combination with the single author rule, fostered competition over collaboration, leading users to focus on authoring rather than reading, rating and improving what others have authored. This probably yielded to needless redundancy, moderate information disclosure, and a moderate level of debate. Thus the power of the community was not fully exploited to improve the quality of the posts. One of the major changes we are considering therefore is to introduce an open authorship rule, such as in Wikipedia, and to rely mostly on intrinsic incentives such as voluntarism and reputation.

Further empirical evaluations will also aim at understanding the effects of using the tool on users' cognitive abilities and at a collective level, on deliberation performance. With respect to the first issue, some studies have provided empirical evidence that argumentation platforms improve learning [24,42] and increase problem solving actions as well as shared knowledge construction in group discussions [4,22]. Overall, there is little doubt that argumentation improves reflection and analysis, but there has been no convincing evidence, in our opinion, that argument based platforms effectively mediate conversations by helping users to create and provide better content. These conclusions were partly confirmed by our study: the outcome of a user satisfaction survey submitted to students after the experiments shows that about 90% of users appreciated the tool's capability to help to organize knowledge, especially in terms of compactness and content visibility. The heavy level

<table>
<thead>
<tr>
<th></th>
<th>Efficiency comparison of raters and non-raters.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Non-rater mean rank</td>
<td>56.45</td>
</tr>
<tr>
<td>Rater mean rank</td>
<td>85.03</td>
</tr>
<tr>
<td>Mann-Whitney U</td>
<td>1185.00</td>
</tr>
<tr>
<td>Significance</td>
<td>.00</td>
</tr>
</tbody>
</table>
of moderation implied however, that people found it difficult, at least at the beginning, to use the argument formalism
properly.

In our next experiments our aim will be to compare the Deliberatorium with sharing tools in particular with forums and
wikis, in supporting effective deliberation. Each tool will be used by separate groups of users in discussions on the same to-
pic. The prevailing incentive type for all user groups will be intrinsic in order to improve the experiment design.

5.2. Implications for design

In the following we report implications of our study on possible ways to improve the organizational and technical design of the tool.

5.2.1. Incentives

User behavior in our study was, we believe, influenced by the incentive structure in place. We relied on an incentive mix
that was clearly biased toward extrinsic incentives due to the awards for the best participants. An indirect extrinsic incentive
was also probably created by the students’ perception of the experiment as a course task for which they would be evaluated
by their professors. The choice to use extrinsic incentives was motivated by the need to create a larger level of activity than a
purely voluntary and open participation might have produced.

This kind of extrinsic motivation may have had a greater positive impact on the quantity of participation than on its qual-
ity. This may explain why students apparently cared more about posting arguments rather than rating them, as shown by the
Pareto analysis above where the “create” event had the largest “top user” group compared to view and rate events. On the
other hand, the emphasis on creation rather than rating could be a result of the shortness of the time window available for the
experiment (3 weeks): it is reasonable to conjecture that in the early phase people would focus on generating new con-
tent and they would move to quality improvement and analysis later on. However we have no means to evaluate the relative
importance of the two effects.

This suggests three lessons. First, user behavior can be strongly influenced and even distorted by available incentives. Sec-
ond, relying on extrinsic incentives can be a way to encourage participation, but there is a risk of inducing significant dis-
tortion in user behavior. Moreover, other studies on peer-to-peer online communities show that intrinsic incentives
definitely play a more important role than extrinsic ones in predicting intensity and quality of participation [38]. This caveat
may not be valid with real experts who instead, could be sensitive to the presence of extrinsic rewards [33]. Third, in the case
of argumentation tools, incentives should be designed to induce desirable argumentation practices and good knowledge
structures; in other words, just evaluating post quality as it happens in forums could not be enough. One possibility is to
evaluate content quality and people’s reputation through more sophisticated belief aggregation algorithms, whose adoption
is possible thanks to the structure of the argument map. We will discuss this issue more in depth in Section 5.2.2.

5.2.2. Organizational structure

Our data suggests some interesting implications for the organizational design of the community. On average, arguments
received only a few ratings. Furthermore, about 50% of posts were viewed ten or fewer times by someone other than the
author of the post or the moderators. On the other hand we observed a rapid increase in the number of ratings over time,
from the 295 votes given during the first week to the 2032 ratings given in the third week. Eventually, the growth in ratings
outpaced the growth in posts.

Those findings indicate that in our experiment exploration and exploitation activities were highly unbalanced in favor of
the former: it seems users were much more concerned with adding knowledge through authoring than in evaluating it. This
is in part due to the experimental setting: (i) in a short time window exploration can be expected to prevail over exploita-
tion; (ii) a non-expert community may tend to produce a shallow debate because of the lack of thorough knowledge; and (iii)
the incentive structure may induce to focus on authoring.

Those distortions however help to see that authoring has to be sustainable, i.e. the community should generate content at
such rate that it can be properly organized, evaluated, and refined. These findings suggest that we might want to introduce
mechanisms that restrain content generation by users, such as a more fine-grained hierarchy based on reputation in which
highly-rated authors have more visibility or authoring power. We might also want to use reputation systems to identify
moderators within the community, perhaps coupled with some random rotation among the best ones to ensure adequate
turnover in the moderators group.

5.2.3. Rating and aggregation systems

The need for reputation and content quality assessments shows that self-governance of an online community is highly
dependent on the functioning of the rating system. Furthermore this aspect is particularly important in a deliberation com-
community where better decisions can be given with an increasing number of reliable ratings.

Though in our tool the rating system was very basic and similar to those of many online discussion forums, our findings
show that even this primitive form of rating is desirable since it seems to induce higher participation levels and better under-
standing of the logic of the systems: raters check their own posts over and over and also contribute to the deliberation by
developing other users’ arguments.
As indicated above in our discussions about incentives, the structure of an argument tree renders the addition of a more sophisticated rating and aggregation algorithm possible. Introne [18] proposes an algorithm based on Dempster–Shafer theory [37] through which top-level posts are rated on the basis of the amount of support coming from the underlying chains of pros and cons. The main advantage of having a belief propagation system is that it enforces users’ understanding of the formalism, supporting the desired behavior (e.g. users will provide good pros to support an idea they like or will attack a con attached to the same idea). The price to pay is that it requires having an argument map whose structure satisfies the requirements of a formal propagation algorithm, which can increase moderation costs.

In an argument map ratings can be elaborated to build indicators of consensus or activity that could be used to monitor the debate, and mediate user attention. Indicators can be enhanced by using visualization tools such as tag clouds or history visualization flows as well [13,48].

6. Conclusions

In this paper we presented the result of our analyses on the use of a new mass collaboration platform called the Deliberatorium. The aim of the Deliberatorium is to support large, geographically dispersed communities of users in the collective deliberation of complex and controversial issues. The key difference between the Deliberatorium and other large-scale collaboration tools such as forums, blogs, chat rooms, and wikis is that it supports the construction of shared knowledge representations in the form of argument maps, with a logic – rather than time-based structure. The Deliberatorium differs from other argumentation tools in that it incorporates features designed to enable deliberations with much larger scale user communities. These features include technologies adopted from other large-scale deliberation tools (including search, history, watchlist, chat and web forums) plus the introduction of a moderator layer and moderation tools supporting knowledge mapping and organization. In this paper we have reported some preliminary results of a first 21 field experiment with a community of about 160 users. To our knowledge the Deliberatorium represents the first argumentation platform to be applied at this scale while the online argument map built by novice users is also one of the largest, developed on a complex topic; given that the experiment was held in a considerably short period of time.

With respect to our main research question, i.e. can argumentation tools effectively mediate on line discussion to support collective deliberation, we obtained contrasting results. The final map showed good coverage of the current debate on biofuels, creating thousands of, on average, pertinent and good quality posts. The collection of such content was very efficient considering that it was produced by a community of novice users in just three weeks. However, the content organization has been possible with an over-reliance on moderators. In other words, we observed that it was difficult for users to switch from the conversational attitude usually present in discussion forums to an even mildly formalized argument-based discussion. When used properly, argument mapping helps to organize and summarize information, but as with other formal knowledge representations, it does not work well for mediating dialog, nor does it support the movement from less structured exploratory dialog to more structured knowledge representations. Partial evidence for this point is the intense use we observed, during our study, of conversational tools like forums and chats for “out of map” communication.

While we find a certain amount of moderation unavoidable, we think that major design efforts are needed to lower moderation costs as much as possible. We envisage two non-exclusive possibilities: (1) crowd-sourcing the moderation so that the costs are spread across a large group of people; (2) improving the knowledge representation to shorten the gap between argument-based dialogues and the way people actually make conversations. Both aspects are the foci of our current research.

Acknowledgements

Luca Iandoli developed part of the research activities as a Fulbright visitor at the Center for Collective Intelligence (Massachusetts Institute of Technology). The Fulbright Program support is gratefully acknowledged. This paper reports some of the research results obtained within the research grant PNII - IDEI 810/2008 - “The Development of Theoretical Fundaments of Knowledge Ecosystems and Their Applications on Economy and Healthcare”, funded by CNCSIS –UEFISCUS. The authors thank the Science Museum of the city of Naples (Città della Scienza) for sponsoring the scholarships used as incentives in the experiment.

References
