Socially-Augmented Argumentation Tools:
Rationale, Design and Evaluation of a Debate Dashboard

Luca Iandoli[a,b], Ivana Quinto[b*], Anna De Liddo[c] and Simon Buckingham Shum[c]

a. School of Systems and Enterprises, Stevens Institute of Technology, Hoboken, NJ, USA
b. Department of Industrial Engineering, University of Naples Federico II, Naples, Italy
c. Knowledge Media Institute, The Open University, Milton Keynes, UK

*(corresponding author)

Professor Luca Iandoli
School of Systems and Enterprises, Stevens Institute of Technology,
Castle Point on Hudson, 07030 Hoboken, NJ (USA)
liandoli@stevens.edu, (001) 201 216 8104
Department of Industrial Engineering, University of Naples Federico II
80, Piazzale Tecchio, Naples, Italy, 80125
iandoli@unina.it, (0039) 081 7682935, Fax (0039) 081 7682154

Dr Ivana Quinto
Department of Industrial Engineering, University of Naples Federico II
80, Piazzale Tecchio,Naples, Italy, 80125
Ivana.quinto@unina.it, (0039) 081 7682932, Fax (0039) 081 7682154

Dr Anna De Liddo
Knowledge Media Institute, The Open University
Walton Hall, Milton Keynes, United Kingdom, MK7 6AA
anna.deliddo@open.ac.uk, (0044) 1908 653591, Fax (0044) 1908 653169

Professor Simon Buckingham Shum
Knowledge Media Institute, The Open University
Walton Hall, Milton Keynes, United Kingdom, MK7 6AA
simon.buckingham.shum@open.ac.uk, (0044) 1908 653591, Fax (0044) 1908 653169

Abstract. Collaborative Computer-Supported Argument Visualization (CCSAV) is a technical methodology that offers support for online collective deliberation over complex dilemmas. As compared with more traditional conversational technologies, like wikis and forums, CCSAV is designed to promote more critical thinking and evidence-based reasoning, by using representations that highlight conceptual relationships between contributions, and through computational analytics that assess the structural integrity of the network. However, to date, CCSAV tools have achieved adoption primarily in small-scale educational contexts, and only to a limited degree in real world applications. We hypothesise that by reifying conversations as logical maps to address the shortcomings of chronological streams, CCSAV tools underestimate the importance of participation and interaction in enhancing collaborative knowledge-building. We argue, therefore, that CCSAV platforms should be *socially augmented* in order to improve their mediation capability. Drawing on Clark and Brennan's influential Common Ground theory, we designed a Debate Dashboard, which augmented a CCSAV tool with a set of widgets that deliver meta-information about participants and the interaction process. An empirical study simulating a moderately sized collective deliberation scenario provides evidence that this experimental version outperformed the control version on a range of indicators, including *usability, mutual understanding, quality of perceived collaboration,* and *accuracy of individual decisions*. No evidence was found that the addition of the Debate Dashboard impeded the quality of the argumentation or the richness of content.

Keywords: Computer-supported argument visualization, Grounding process, Common Ground, Debate Dashboard, Collective deliberation, Visual feedback

1. Supporting collective deliberation through socially-augmented knowledge mapping tools

Computer-supported argument visualization (CSAV) platforms assist their users in identifying, structuring, and settling issues using argument maps (Buckingham Shum, 2003). An argument map is a visual representation of the informal logical structure of an argument (Walton, 2008). Depending on the representational scheme, it displays the constituent elements of the argument (such as issues, claims, premises, and evidence) as a tree or network, with nodes in the network expressing the elements, and arrows expressing key relationships, such as *evidential support and challenge* (van Gelder, 2007) or the underlying argumentation scheme, such as *argument by analogy, or argument by authority* (Reed and Rowe, 2004; Walton, *et al.*, 2008; Buckingham Shum and Okada, 2008).

A wide range of representational schemes has been devised within different research communities (e.g. law; design; philosophy). Computational argumentation research has developed more formal logic and mathematical models with an interest in reasoning over the model in order to evaluate claims or prove properties automatically (e.g. Rahwan and Simari, 2009). However, there is a tradeoff between computational power and usability: argument models that a computer can evaluate require more careful construction, while semiformal models primarily designed to aid human reasoning make fewer assumptions about the formal correctness of the structures. CSAV tools from the latter class have been applied in the context of education and learning (Andriessen *et al.*, 2003, Suthers 1995; Okada and Buckingham Shum 2008; Scheur *et al.*, 2011), design rationale capture (Moran and Carroll, 2006; Carroll, 2012), group deliberation (Gurkan *et al.*, 2010), individual or collective decision-making support systems (Karakapilidis *et al.*, 2009), and joint problem solving (Cho and Jonassen, 2002; Lu, 2008).

Beyond the cognitive affordances of diagrammatic representation, a semiformal representation is more computationally tractable than a natural language prose expression of an argument in a discussion forum. For instance, through the implementation of a computational engine, it is possible

to aggregate the amount of belief that different users assign to pros and cons, and rank alternative positions in terms of their rational support rather than through traditional online voting (Introne, 2009). Network and semantic analytics can be generated to show the contributions that protagonists make (De Liddo, *et al.*, 2011), and metrics can be defined to evaluate the quality of the deliberation (Klein, 2012).

Critical to the research question that this paper addresses is evidence from an approach that has been relatively successful in real world settings, based on the work of Rittel and Webber (1973). Their Issue-Based Information System (IBIS) represents a deliberation in terms of issues to be solved, positions addressing them, and associated tradeoffs expressed as pros and cons. A graphical hypertext rendering of IBIS (Conklin and Begeman, 1988) has since been implemented in many software tools. Conklin (2003), Buckingham Shum and Sierhuis (2008) and Culmsee and Awati (2011) document case studies in which this kind of mapping added significant value to collective deliberation. However, a key ingredient in making this approach work is the role of an expert facilitator, who assists synchronous deliberation (face-to-face or online) by mapping speakers' contributions, making many decisions about how to record and structure the flow of conversation in an argument map in real time. Our prior research has studied in detail the mapper's initial learning curve (Buckingham Shum, *et al.* 1991) as well as expert fluency in such techniques (Conklin, 2006; Selvin and Buckingham Shum, 2012). The facilitator plays a crucial role in coordinating the verbal conversation around the evolving map: participants thus receive significant support to ensure that the map augments, rather than disrupts, coordination and interaction. The question is whether CSAV can be successful when migrated online for asynchronous interaction, and scaled up beyond a roomful of people, and moreover, deployed without facilitators.

This is the focus of research and development into Collaborative CSAV (CCSAV): mediating large-scale deliberation among many participants. In comparison to traditional conversational tools such as forums or blogs, which are organized around a chronological timeline, CCSAV is intended to draw more attention to logical relationships, through the rational organization of users' contributions into maps. The most convincing examples of CCSAV platforms in large scale use by untrained users come from e-democracy initiatives to engage citizens in public debate. IBIS has been deployed in such initiatives, where it has been used to varying degrees of proficiency— but these examples still remain relatively sparse (Iandoli *et al.,* 2009; Debategraph[1]). In formal educational contexts, CCSAV tools have not reached widespread adoption beyond small scale team working. One clear reason for this is that argumentation tools require users to learn a kind of 'grammar' to translate their thinking into structured rather than prose arguments, whereas there is evidence that, on average, people do not exhibit good argumentation skills (Kuhn, 1991), and they can become proficient with and reap the benefits of argument-based representations only after systematic and repeated training (Twardy, 2004). As discussed elsewhere (Buckingham Shum, *et al.*, 1991), these tools were designed specifically to foster greater reflection, but the greater cognitive effort required naturally trades-off against immediate learnability.

The hypothesis that we test in this paper is that compared to the dominant, more social, but less structured platforms, an adoption obstacle for CCSAV implementations to date is the lack of attention to smoothly supporting users' interpersonal coordination and interaction. Mapping tools disrupt conversations by design: chronological structure is in the background, in order to foreground logical structure as networks of conceptual connections. We posit that by favouring the

---

[1] *Debategraph* exemplifies the current state of the art in IBIS mapping tools used by the general public for deliberation over societal dilemmas: http://debategraph.org

objectification of conversations into knowledge maps, CCSAV tends to neglect the role of interaction among participants in the collaborative creation of knowledge. While social interaction is often disparaged as aimless and distracting, it can in fact play a very important role: (i) it helps strangers to get to know each other better; (ii) it makes a conversation more engaging; and (iii) it provides a way for participants to exchange meta-information about the object of their conversation, in other words, to talk about what they are talking about. In this paper we offer an argument, and empirical evidence, to show that CCSAV tools can be designed to deliver both knowledge mapping support and socially salient meta-information about the participants and the interaction process, while preserving their ability to support artefact-centred discussions. We hypothesise that socially augmented CCSAV will be able to retain the advantages that are traditionally associated with argumentation tools as well as to improve the platform's usability, deliberation process and outcome.

In this paper we present a new CCSAV tool, called the *Debate Dashboard*, built through the integration of a mapping tool, *Cohere* (De Liddo and Buckingham Shum, 2010), and a set of user interface 'widgets' that mediate and reflect back additional metadata about social and conversational interaction. We report empirical evidence from an experiment in which we compare the performance of the users of this 'socially augmented' CCSAV tool, with that of a control group using a non-augmented tool. Our results show that users of the augmented platform outperformed users of the non-augmented mapping tool on several indicators including *mutual understanding, perceived quality of collaboration, accuracy of individual decision,content production* and *domain coverage*.

The paper is structured as follows: in §2 we review the existing literature on online argumentation and discuss the critical issues that arise when argumentation platforms are used to mediate online debate. We then (§3) describe our theoretical framework, which is based on Common Ground theory (Clark and Brennan, 1991); and in §4, report on how this was operationalised as a set of features in the Debate Dashboard. In §5, we present the design of a field test with an online community of 64 participants, tasked with predicting commodity price trends based on their analysis of information sources. §6 presents the results of our experiment, which are discussed in detail in §7, before we draw conclusions and identify directions for future research (§8).

2. Online argument mapping tools for Collective Deliberation

The etymology of *argumentation* lies in the Latin word *arguere*, which means to clarify, emphasize, or demonstrate the reasonableness of a position, but also to debate, discuss and persuade. The *ability to argue* is an essential whenever people must defend assertions or actions, or react to opinions from others (Kuhn, 1991).

Argumentation theorists have focused on the identification of formalisms, rules and principles that can guide people to the creation of well-crafted arguments, including methods for visual representation of an argument's components (Wigmore, 1913; Toulmin, 1958). Different approaches to argument representation and scaffolding have been developed, which differ mainly in terms of the adopted argument ontology. In general, such analytical approaches, by require their users to articulate opinions using a limited set of constituent elements, and are supposed to enable evidence-based, coherent reasoning, offering visualizations of the logic that lies behind the reasoning.

During the last two decades, researchers and designers have created computer-supported argumentation systems to provide users with interfaces that allow them to create, edit and navigate an argument map. The field of artificial intelligence has sought to model the formal properties of arguments and develop end-user applications that benefit from such reasoning (Bench-Capon and Dunne, 2007; Rahwan and Simari, 2009). Argument maps enable users to visualize concepts and

the inferential relationships among these concepts; they also allow for the inclusion of additional contents (e.g. annotations) and other knowledge resources (e.g. websites). Shared visual representations offer opportunities for focusing collective attention, envisaging new scenarios, coordinating actions, and potentially improving the comprehension and retention of knowledge (Okada *et al.,* 2008).

Although several studies have detailed the desirable effects that should be produced by computer-based collective argument maps — such as (i) representational guidance (Suthers *et al.,* 2003), (ii) support to participants to clarify their thinking (Brna *et al.,* 2001), (iii) make this thinking visible to others (Bell, 1997), (iv) foster information and knowledge awareness (Englemann and Hesse, 2010) — in the literature to date there is little unmitigated support for these advantages; rather the results are mixed and unclear (Scheuer *et al.*, 2010)

That said, diverse motivations exist to justify the use of CCSAV tools to support group deliberation. First, collective discourses and debates can help group members to reason better (Mercier and Sperber 2011), so argument-based platforms should encourage criticisms and comparison of diverse, alternative points of view. Second, the argument-based format it is supposed to favour critical thinking (Twardy, 2004; van Gelder, 2007) and evidence-based reasoning (Carr, 2003). Third, large groups could use these tools to systematically and comprehensively explore and map an extensive debate on a selected discussion topic (Gurkan *et al.*, 2010). Finally, it is expected that CCSAV can facilitate contents localization and, unlike time-centric tools, such as online forums, discourage information cocoons and flame/edit wars (Klein, 2010).

Nevertheless, despite these expected advantages, argumentation technology is not widely used in organizations or online communities. CCSAV typically requires users to undergo intensive training to become proficient with the formalism (Twardy, 2004; van Gelder, 2003); in the organizational setting, this requires strong internal sponsorship coupled with individual commitment (Conklin, 2006). There are high coordination and moderation costs when the deliberation involves many users (Gurkan *et al.*, 2010).

While the lack of adequate argumentation skills, together with steep training and supervision costs can contribute to make CCSAV's value proposition less attractive, in this paper we argue that an even more fundamental obstacle for users can be identified in the difficulties they experience using a formal artifact, such as a knowledge map, to mediate dialogue. In particular, argument-based graphical representations can appear unnatural and unintuitive, when compared to other more familiar ways of coordinating and organizing interaction, such as conversations (Wikes-Gibb and Clark, 1992). CCSAV tools prioritize formal representation of contents over temporal flow, and they do not usually exhibit the turn-taking structure that is typical in conversations, making collaboration more awkward, without a clear and immediate visible benefit.

However, despite this critical limitation, research on online CCSAV has focused mainly on knowledge representation issues, while neglecting or underestimating the role of social and conversational processes surfacing in online interaction. Therefore, the aim of this paper is to investigate how CCSAV mediation capability can be enhanced by implementing affordances for social and conversational interaction, and to examine the extent to which social augmentation may improve users' experience and task performance. With this in mind, the central research question we explore in this work is:

RQ:*Can we retain the advantages of online collaborative argument mapping tools while improving their ability to mediate social interaction taking place in the virtual environment? More specifically:*

- **RQa***: Is it possible to make a collaborative argument mapping tool more conversational, social and people-oriented in order to improve the performance of distributed collaboration?*

- **RQb:** *Can such a platform be used in concrete tasks such as collective deliberation to enhance group and individual performance regarding both the quality and the output of the collaboration process?*

3. Theoretical Foundations

3.1 Participation and knowledge reification in the creation of shared knowledge

According to Wenger (1998), knowledge creation in communities of practice requires the interaction of two crucial and constituent processes, namely *participation* and *reification*. *Participation* refers to the process of being active in ongoing collective action, by interacting with other participants and sharing ideas and information, but also common goals, values, resources, workload, and means for action. *Reification*is a process by which an abstract concept is objectified into something that can be seen, used, operated upon or manipulated by other participants for various purposes (communication, understanding, learning, negotiation, etc.). It is the process of making an idea into an artefact that can be more shared with others. Examples of reification are bylaws, organizational charts, contracts, quality manuals and Wikipedia pages. The artefacts we are interested in this paper are created by participants, in a grass-roots manner.

According to Wenger, in a community of practice these two processes are mutually reinforcing and must be properly balanced. In particular, participation can serve the critical function of exchanging meta-knowledge useful to assess, criticise and redesign shared knowledge embodied into artefacts.

3.2 Artefact-centered discourse and the creation of common ground

CCSAV tends to rely on reification andtakes participation for granted. Nevertheless, even when participants find it easy to talk about a map, they typically would find hard to exchange this type of meta-information "through the map". The objectification of conversations into argument maps obscures another range of critical meta-information. According to Clark and Brennan (1991), in a conversation, participants not only exchange content, but also meta-information that isneeded to ascertain whether or not they understand each other.Usually this information is conveyed by verbal and nonverbal communication acts. Feedback of this sort can be used as evidence of mutual understanding and mutual knowledge. The construction of this form of mutual knowledge is called *grounding*.

When the conversation is mediated by any kind of communication technology, part of the conversational feedback that is available in face-to-face discussions is either missing, or has to be provided with some extra effort. For example, in a face-to-face conversation, people can use facial and body expression to express agreement or to show their attention, while in computer-mediated conversations participants may provide this information by using different techniques, for instance, by typing something, or, in particular, using emoticons. The theory of common ground posits that mediated communication is always less efficient than face-to-face interaction, in which a number of communication constraints help to reduce the *grounding cost*, i.e. the effort needed to build mutual understanding (Table 1).

Table 1. Constraints in communication media

Constraint	Clark *et al.'s* definition	Supported by common CCSAV	Definition adapted to online conversations
Audibility	Participants hear each other as well as other sounds in the physical environment	No	Not Applicable
Co-presence	Participants share the same physical space	No	Participants are mutually aware that they share a virtual environment
Co-temporality	B receives words at roughly the same time as A produces	No	A participant can receive several messages at roughly the same time as they are produced by other participants
Mobility	Users can move around physically	No	People can move around in a shared virtual space
Reviewability	B can review A's message	Yes	Messages persist and can be reviewed
Revisability	B can revise a message to be sent to B	Yes	Message can be revised before being sent
Simultaneity	A and B can send and receive at once and simultaneously.	No	Not Applicable
Sequentiality	A's and B's turns cannot get out of sequence.	No	Participants can identify the reply structure
Tangibility	Participants can touch other people and objects in the physical environment	No	Not Applicable
Visibility	A and B are visible to each other	No	Participants see the actions of others or have evidence of actions performed by other users in the shared virtual environment

When fewer constraints are satisfied the costof building common ground is higher. Applying Clark and Brennan's framework to CCSAV tools, it is straightforward to verify that the grounding cost of argument-based communication is very high.Eight out of ten constraints are not satisfied, namely *co-presence, audibility, visibility, tangibility, mobility, co-temporality, simultaneity* and *sequentiality* (Table 1).

Wenger's theory of collective learning and Clark and Brennan's theory of grounding costs point to potential improvements in the design of CCSAV tools that would augment their mediation capability. Users need additional ways to exchange and access socially salient meta-information in order to achieve a better balance between knowledge objectification and participation.This could be achieved by making participants and their activities visible to one another. Suitably augmented CCSAV tools could deliver a form of conversational feedback that would reduce grounding costs.

3.3 Suggested types of feedback

Our hypothesisis that to augment mediation capabilities of CCSAV tools we need to fill a meta-informational gap. We can identify three general categories of meta-information that are natively missing or not systematically delivered by currently available CCSAV platforms:

- **Community (*who*):**Community-level meta-information is aimed at making users more knowledgeable about the social landscape of the virtual community;
- **Interaction process (*how*):**This type of meta-information is meant to support the grounding process and facilitate the growth of mutual understanding;

- **Knowledge absorption (*what, where*):** Content-related meta-information should facilitate users in making sense of online discussion and identifying points in the debate where their contributions could be more effectively placed.

Each of these types of meta-information can be delivered to the users by providing them with specific forms of feedbackthat is generated as a by-product of their online engagement. This feedback serves the following purposes:

i)   Making social engagement easier and more "affordable";
ii)   Favouring the construction of mutual understanding and the accumulation of shared knowledge;
iii)   Making the virtual platform socially translucent (Erikson and Kellogg, 2002; McDonald *et al*., 2009), by allowingusers to display an online identity, and helping them develop awareness of other participants' activities.

Figure 1 shows a more detailed classification of feedback types. Some of these feedbacktypesareadapted directly from Clark and Brennan's original formulation(Table 1).Below, we will describe each category of feedback in more detail. Each feedback type points to specific user requirementsthat we address in the design of the Debate Dashboard.

Community feedback. According to Social Translucence theory (Erickson and Kellogg, 2002), making socially salient information visible enhancesparticipants' awareness and accountability, and provides resources for structuring and fostering interactions. Additional studies provide evidence that better knowledge of the social landscape in a virtual community will increase the level of participation, strengthen the flow of information and knowledge, intensify cooperation as well as the level of commitment to common goals, members' engagement in community activities and the level of satisfaction about group effort (Bruffee, 1993; Dede, 1996; Erickson and Kellogg, 2002; Royal and Rossi, 1996; Wellman, 1999).
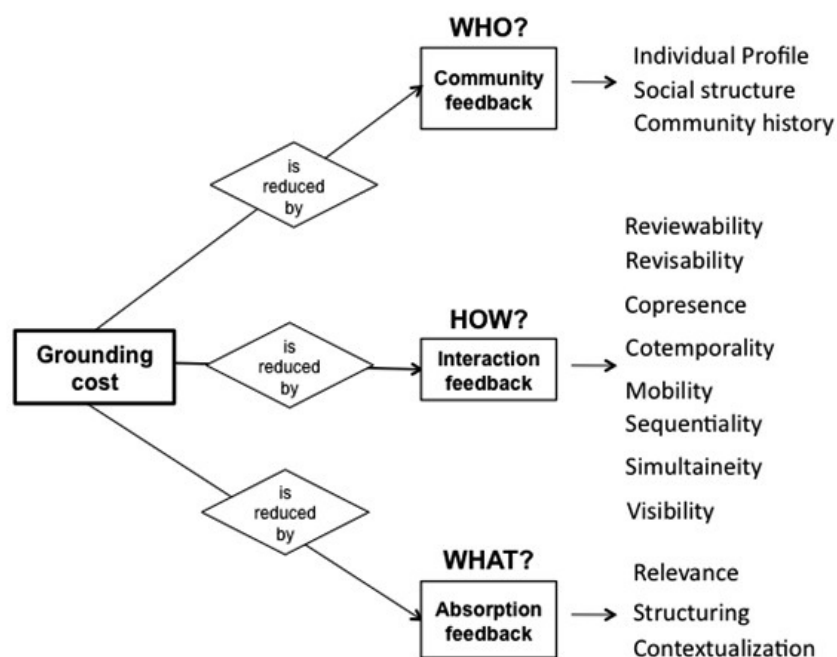

Figure 1. Model of suggested feedback. A more detailed classification of feedback types that should augment mediation capabilities of CCSAV tools

To implement community feedback inthe Debate Dashboard, we identify the following requirements:

- **Individual Profiles:** Providing members' personal information, such as name, age, place of birth, job/occupation and hobbies is common in social media, to promote better mutual acquaintance among members, and the creation of online identities.
- **Social structure:**Adopting the traditional graphic formalism used in Social Network Analysis, participants are represented as nodes connected by links that are inferred in real time by the system from the conversational exchanges taking place between any two members. This visualization helps members to develop a better understanding of the social dynamics in the community; for instance, participants can know who is speaking to whom and the intensity of this relationship as measured through the frequency of exchange. Additionally, users can gain insight into the structure of the network, e.g. ascertaining who the most central individuals are, or which members fill structural holes, bridging two or more separate subgroups.
- **Community history:** Information about activity and participation levelscan be delivered through time-based, cumulative view of the community's past activities (e.g. most recent, most viewed, or top rated).

Interaction feedback. With the exception of *Tangibility, Audibility, and Simultaneity,* which are not applicable in the virtual asynchronous settingthat is typical of CCSAVs, interaction feedback favouring the construction of common ground can be provided to satisfy all of the other constraints:

- **Copresence:** Users want to know who is online and be aware that the shared space is active. Knowing who is online should make easier to start a conversation, by favouring opportunities for online "rendezvous". When people "meet" virtually they feel they are not alone; they know someone is available for communication, and have the feeling of being listened to.
- **Cotemporality:** As we are focusing on asynchronous conversations, participants are not guaranteed to receive instantaneous replies to their contributions. Delays make the construction of common ground more difficult. Nevertheless, a partial form of cotemporality is established by alerting people when a new contribution has been made, reducing the time lag from production to reception. RSS-like feeds and other forms of notification, such as the "follow" function in Twitter, are possible ways to provide the users with effectively real time alerts about reactions from other participants.
- **Mobility:** It should be possible to visualise users' "movements" across different topics or discussion groups.
- **Sequentiality:** In an argument-based platform, the representational focus is on the logical rather than the chronological structure of a conversation and users' contributions are displayed in a map rather than along a timeline. The absence of sequentiality is arguably a major disruption to smooth interaction, because speakers do not have immediate evidence about hearers' understanding of their utterances, and cannot fix misunderstanding quickly. One way to overcome the problem is to provide users with a parallel discourse representation that is able to trace the reply structure.
- **Visibility:** An easy way to provide feedback on what other members have been doing is to make available key statistics about users' activity and history, such as the number of posts and connections created about a specific topic. In this way, users' past actions are rendered more visible to the other community members.
- **Reviewability and Revisability:** These features are typically available in CCSAV platforms: users can review posts published by other participants at any time, and revise their own posts before and after publication.

Absorption feedback. When the number of people using a CCSAV platform increases, the amount of "boxes and arrows" usually grows as well, and argument maps become cluttered and hard-to-read. So-called "spaghetti" images of unfiltered hypertext networks (Hair, 1991; Loui et al*.,* 1997)impede sense-making. In order to cope with this problem, we introduce a class of feedback that we call *absorptionfeedback*, after the concept of *absorptive capacity* introduced by Cohen and Levinthal (1990). Unlike conversational feedback, which is designed to smooth interpersonal interaction, absorption feedback aims to facilitate interaction between individuals and a knowledge artefact representing a discourse structure. In face-to-face conversation, absorption is less relevant because the number of participants is typically small and live interaction makes available a number of ways to summarise the contents of a conversation. The sub-classes of feedback we have identified in this category are:

- **Relevance:** By segmenting the accumulated information into manageable pieces (e.g. topic clusters), this form of feedback is expected to help users explore a larger amount of information, locate relevant content, and find their way into the conversation in the right place.
- **Structuring:** this form of feedback exposes connections between different chunks of information. In this way, people can find trends, patterns or structures in a larger amount of content,and possibly discover new and unanticipated connections.
- **Contextualization:** this form of feedback should help users to establish the context of contributions, reducing the gap between new information and potentially relevant, pre-existing knowledge.

4. The Debate Dashboard

We have argued for the need to provide users with specific types of social and conversational feedback that are not readily available in CCSAV-mediated discussions. In order to validate this claim, we developed an augmented CCSAV tool called the Debate Dashboard that runs on top of an existing online deliberation tool called *Cohere*. Cohere is a web-based asynchronous discussion tool.It combines some of the common features of argument mapping tools with social features, such as social network visualizations and collaborative web annotation (Buckingham Shum, 2008; De Liddo et al., 2011). Some of Cohere's features were not of interest in this study, and were removed from the two test versions.Cohere was extended to provide a significant subset of the meta-informational feedback extensions introduced above.

In Cohere, the online discussion is represented through a deliberation map consisting of nodes and links, in which each post represents a node and each link expresses the connection between posts. Connections have associated labels such as *responds to* question, *supports/challenges* an idea, etc. The structure of the map is based on the IBIS grammar (Rittel and Webber, 1973) that is used in other deliberation platforms (Buckingham Shum, 2003; Conklin, 2003; Gurkan et al., 2010).

4.1 Community feedback

Profile. *Profile* feedback, provided through Cohere's People Page, aims to provide information on "who's who." A user's profile page helps other users find out more about who a user is, and how (s)he contributed to the discussion.Information about ideas created, connections identified, websites annotated, and group participation isprovided in sub-tabson the user's profile page (see *Ideas, , Connections, Websites,* and *People & Groups* tabs in Fig. 2).

Figure 2. User's Profile: The user profile component of the Debate Dashboardprovides historical and socially-relevant information about a community member

Social structure. *Social Structure feedback*isaccessible through the social network tab, and is delivered through a sociogram showing the users contributing to the discussion (Fig. 3). An algorithm calculates each users' degree centrality, a well-known social network analytic reflecting how many ties a node has (Borgatti *et al*., 2005).In the social network visualization, different colours are associated with different degrees of centrality ranging from dark pink (most connected), to highly connected (light pink), moderately connected (turquoise) and slightly connected users (grey).
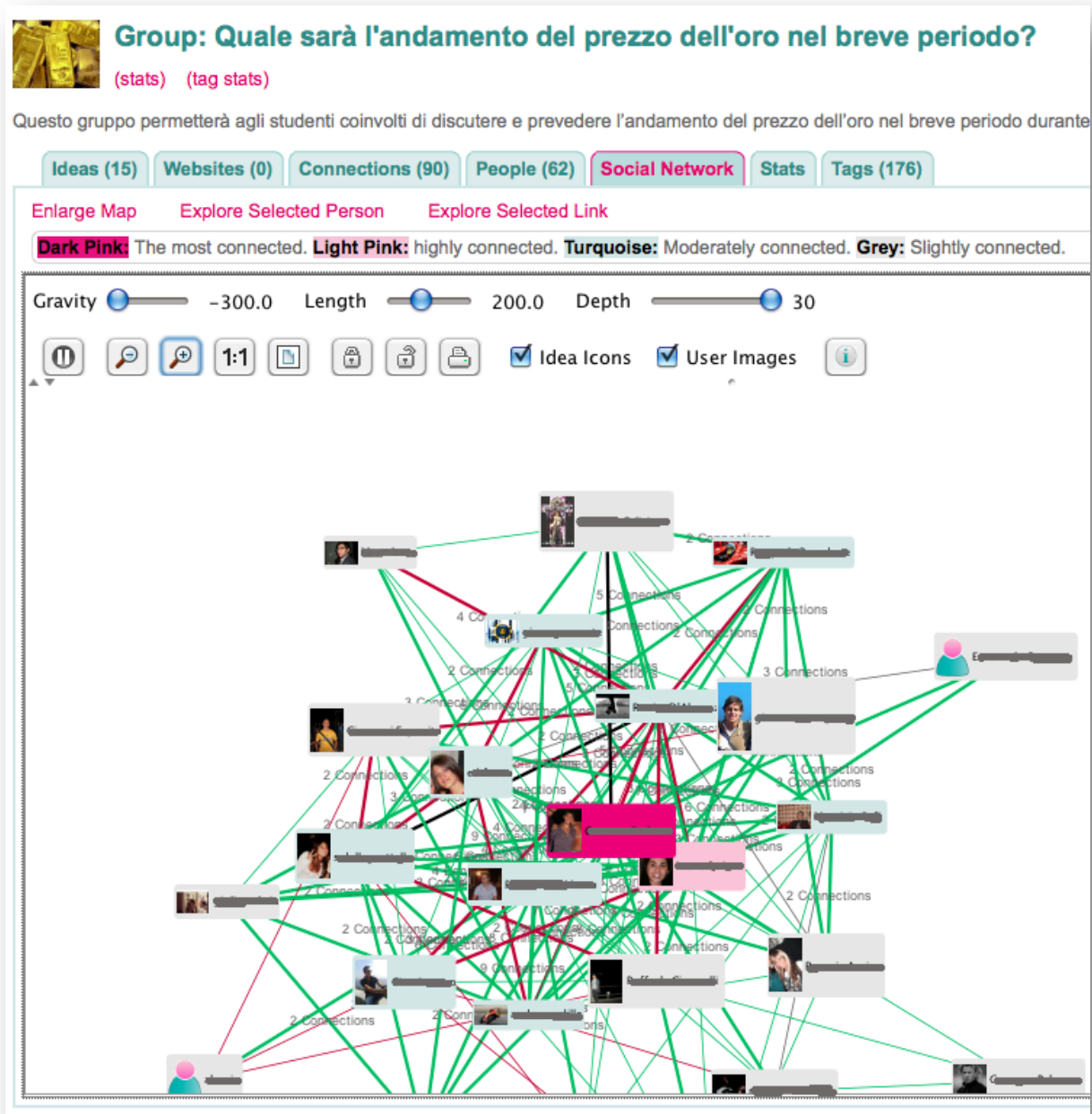
Figure 3. Social network visualization (users' names obscured for privacy)

A difference with traditional sociograms is that the systemalso provides feedback on the *polarity* of the relationship between two users (De Liddo, *et al.*, 2011). An algorithm assesses whether the relationship between two users can be classified as *positive, negative* or *neutral,* based of how many times they have agreed/disagreed or neutrally addressed each other. To provide this feedback, the links between the users in the sociogram are colour-coded as green (agree), red (disagree), or grey (neutral).

Community history.*Community history* feedback traces the record of activities of participants through simplestatistics on the generated ideas.This information is accessed through the *Ideas* sub-tab, under the discussion group *Stats* tab (Fig. 4). The *Ideas* stats page provides the list of generated ideas ordered by three main variables: time (most recent ideas), popularity (most voted ideas) and centrality (most connected ideas).
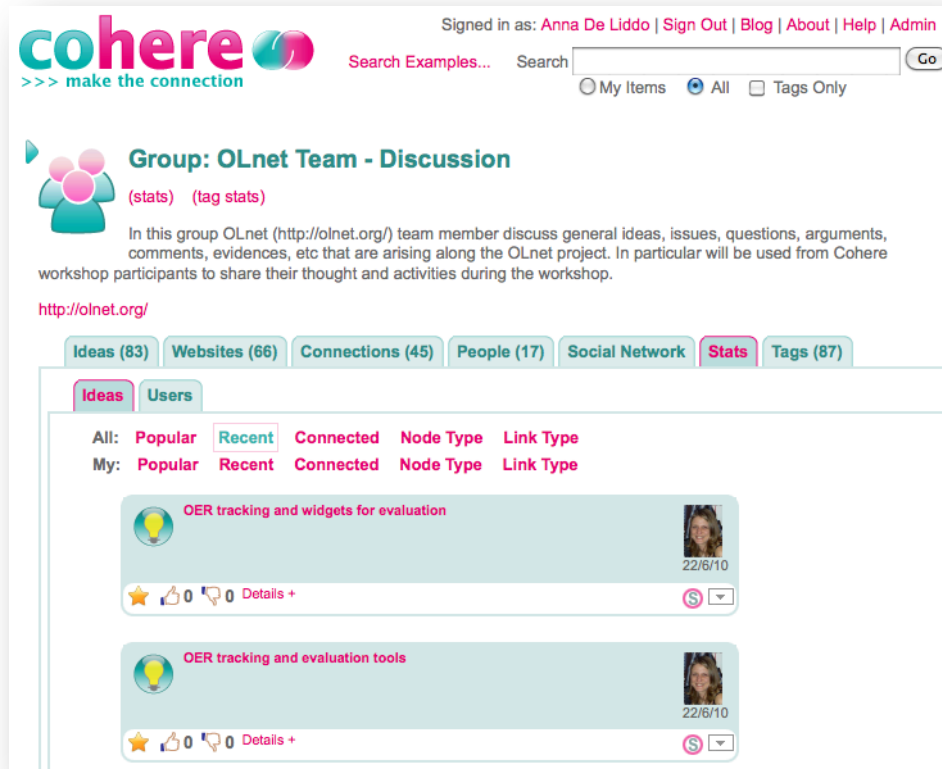
Figure 4. Community history: Community ideas (nodes) can be sorted by different criteria

Recent ideas generates in a timeline view of contributions, analogous to a list of status updates in social media. The most *popular* filter gives immediate feedback on community trends. In order to measure popularity, an algorithm computes the sum of absolute votes (positive and negative votes) for each idea;to break ties, it pushes to the most debated ideas to the top of the list. The most *connected* ideas list provides a list of the 20 most important ideas,as determined by calculating their *degree centrality in the deliberation map.*

The "Node type" and "Link Type" statistics provide a list of node types and link types ordered by frequency of use. This list may help to build insight into the *type* of conversation and its *stage*. For example, a list in which there are more questions than answers may suggest that the topics under discussion are still relatively fluid and the process is in its early stages (brainstorming and ideas clarification by "question making"); on the contrary, a discussion in which there are many pros and cons would rather refer to a deliberation process which is more mature and a discussion group which is actively engaged in exploring controversial issues.

Interaction feedback

Out of the ten constraints identified by Clark and Brennan (Table 1) we discarded *Audibility* and *Tangibility* because those are usually not available or expected in asynchronous online discussions. We decided not to provide feedback about *Mobility*, *Co-temporality* and *Simultaneity*:given the asynchronous nature of Cohere, supporting these features would have required deep changes to the code that would have been too costly and risky. *Sequentiality* is also absent because the deliberation map, by design, organises contributions according to their logical role in the discourse instead of a chronologicaltimeline.As described earlier, *Reviewability* and *Revisability* are readily available in a deliberation map. Finally we added support for*Co-presence* and *Visibility* as describednext.
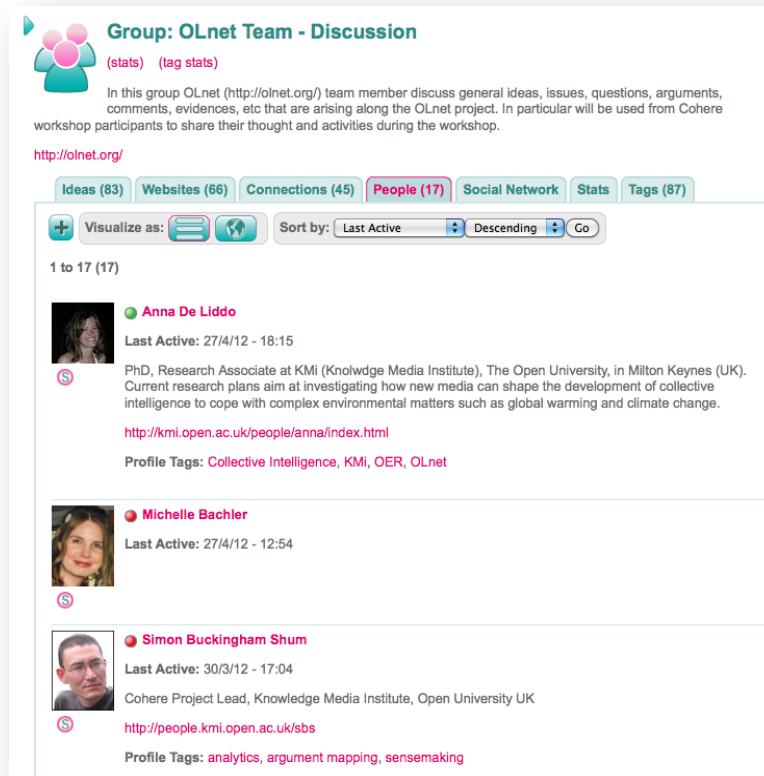
Figure 5. Online presence indicators in the *Cohere* platform: The "People" tab provides the *Co-presence* feature described in Clark and Brennan's Common Ground framework

Co-presence. The *co-presence* feedback shows who is currently online. In each discussion group, by clicking on a tab called "People" it is possible to access the list of users in a discussion group (Fig. 5). As in many other platforms, a green or red circle adjacent to each user's name shows the online or off-line status of a user.

Visibility. *Visibility* feedback shows what a given user has contributed to a given discussion. It is delivered through a Group Stats tab in each discussion group page (Fig. 6). The *Stats* tab presents two main types of statistics for the Discussion group: statistics on users, and on generated ideas. The *Users* stats tab describes each users' performance in terms of three variables: content creation (top node builders), connection making (top connection builders) and socialization (most connected users).
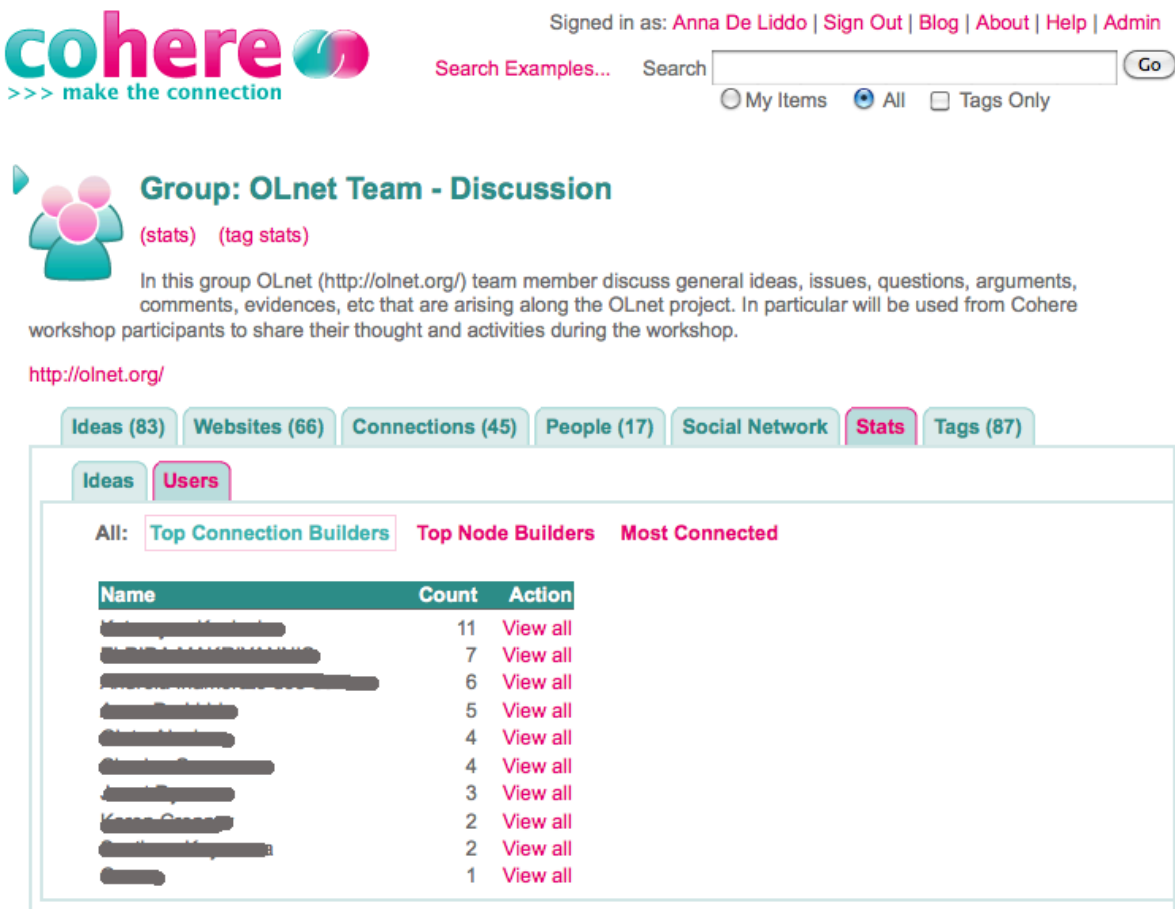
Figure 6. User activity statistics in the *Cohere* platform: The "Group Stats" pageprovides the *Visibility* feature from Clark and Brennan's Common Ground framework (users names obscured for privacy)

4.2 Knowledge absorption feedback

Structuring.*Structuring* feedbackaims to provide information on the output of the conversation.In our case, the main output is the deliberation map built by the group (Fig. 7). The visualization of the map provides users with a bird's eye view of the debate, quickly showing hubs and possibly gaps in the conversation. In order to visualise the map, the users have to click on the "connections" tab in the group discussion.
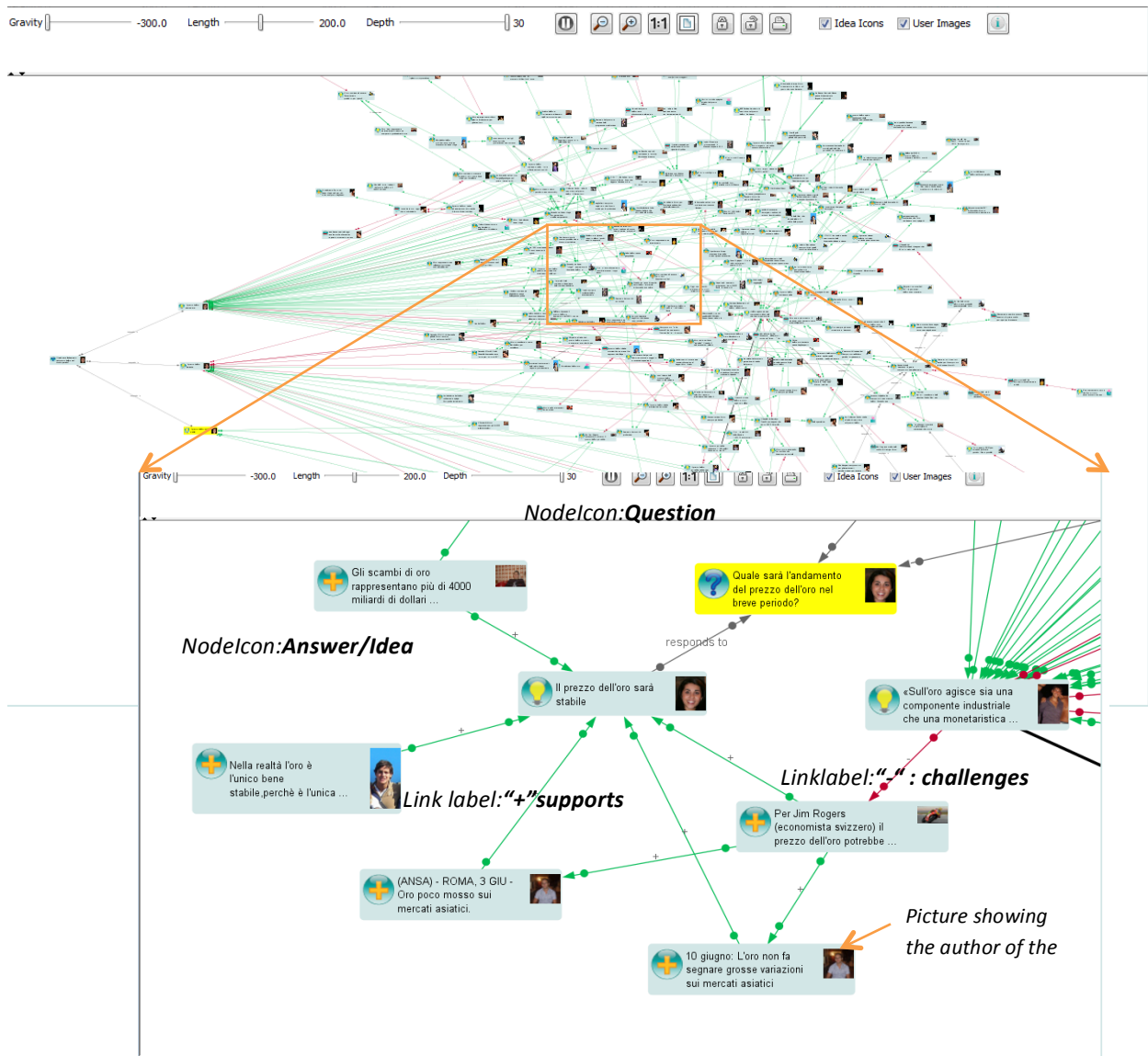
Figure 7. An excerpt of the deliberation map

The deliberation mapis an argument map that is collaboratively constructed throughout the discussion. In particular, it follows the IBIS argumentation formalism: users label their posts as either questions, answers, pros or cons, and each post type will then be displayed with different icons in the map. Concentration of green/red links around an answer provides a visual sense of how well supported or challenged a contribution is, albeit only in terms of the number of connections.

Relevance &Contextualization.*Relevance* and *Contextualization* feedback provides additional graphical clues on the contents developed during the discussion.These features are implemented using Tag Clouds. Users can assign tags when they add posts. Tags are then used to aggregate and retrieve contents by topic (*relevance* feedback), or as a navigational trail (*contextualization* feedback). Users can click on the *Tags* tab in the discussion group (Fig. 8) to see a Tag Cloud, with font size proportional to the frequency of occurrence among the posts.

Figure 8. Relevance feedback delivered through a tag cloud

A "Top 50" tag cloud is displayed in the left sidebar of the discussion group (Fig. 9). Users can navigate the website by clicking on a specific tag, and thus filtering the content in a new result page.

Figure 9. Contextualization feedback

5. Empirical Evaluation

A field-test was conducted with an online community of students. The aim of the field test was to evaluate the effect of social augmentation on collaboration and deliberation outcomes by comparing a control version of Cohere with its socially augmented counterpart, the Debate Dashboard. Specifically, we tested the following hypotheses:

> *H1: Debate Dashboard users will develop greater **mutual understanding** than users in the control group.*

> *H2: Debate Dashboard users will perceive a higher **quality of collaboration** than users in the control group.*

> *H3: Debate Dashboard users will make more **accurate individual decisions** than users in the control group.*

> *H4: Debate Dashboard users will find the platform more **usable** than users in the control group.*

The hypotheses can be justified on the basis of Common Ground theory and Wenger's analysis of the importance of balancing participation and reification. The availability of meta-information through social and conversational feedbacks is supposed to reduce users' burden in the construction of common ground. The reduction of grounding costs is expected to improve users' experience (H4) and users' perception of the quality of the collaboration process (H2). Overall, testing these hypotheses will shed light on whether or not the new social features make participation easier.

5.1 Subjects and Domain Task

An online community of 123 subjects was involved in a single-factor, asynchronous and distributed deliberation field test.. The subjects were students in the same class from a graduate program in Engineering Management, age 19-22, 60% male. Participation was voluntary and participants were compensated with extra academic credits.

The selection of an appropriate domain task was based on the following criteria:

- *Realism*: We wanted to test the platform in a real world decision task. While empirical evaluations are abundant in the CCSAV literature, many studies are designed as short term lab experiments in which users work on abstract tasks (e.g. based on logic) or educational applications (e.g. understanding/analyzing an essay);
- *Task difficulty*: It was not feasible at this stage to involve professional domain experts. However, we had to make sure that our subjects had enough background and skills to make at least educated guesses.
- *Measurability of the outcome*: Unlike other CCSAV evaluations that have been performed on wicked problems for which there is no knowable correct solution, we sought a task with a correct solution to evaluate accuracy of individual guesses.

Taking into account the above criteria, we set up a two-week asynchronous discussion in which participants were asked to predict the price trend of two commoditiesover the short term (Gold and Crude Oil). This task is *realistic* and relevant for economists and business operators. In fact, given that Oil and Gold prices are affected by many uncertain variables, they can be extremely volatile and their forecast is a notoriously difficult task. However, the task cannot be classified as a wicked problem, because there is a unique correct solution, but it is a non-trivial task that could benefit from collective exploration of a large and uncertain decision space.

With respect to appropriate *task difficulty*, our available subject pool consisted of students attending an undergraduate course on economics. Although they were not professional analysts, they had some knowledge about the theory of market equilibrium, and the gold and oil markets had been introduced as course special topics.

The task provided *measurability of decision outcome*, because actual oil and gold prices can be measured and are available ex post for comparison with the subjects' forecasts.

Since students belonged to the same class, we assigned a different case study to each group to prevent inter-group information spill over, and limit personal collaboration outside the two virtual environments. The two tasks are similar in terms of nature and level of difficulty. In the data analysis we will describe in detail a methodology to compare the maps and will show that the two maps are comparable despite the difference in topic.

5.2 Field test design

Since the objective of this study was to compare the performance of two different versions of the CCSAV tool (augmented *versus* plain version), we adopted a one factor, between-subjects experimental design to confront the treatment group (Debate Dashboard) with the control group (Cohere). In order to avoid undesirable influence of pre-existing differences between subjects we performed a random assignment of subjects to the two conditions and performed statistical tests to check that the two groups were not significantly different with respect to demographics (gender, age) and academic performance; the results showed that the difference between two groups are not significant ($t = 1.218$).

5.3 Application and analytics infrastructure

In order to accommodate the requirements of the experimentation design, the two versions were hosted on different servers, with some features available in the full version of Cohere being removed in order to differentiate thisas the control version., Cohere did not provide the real-time tracking logs from user activities needed in order to perform some of the measurements required by our analysis. To save the costly coding of such tracking functionality within Cohere, we adopted the strategy of Pardo and Kloos (2011), using their virtual machine analytics infrastructure. Participants installedthe virtual machine on their computers, and conducted all their sessions inside it, generating activity traces which were uploaded to a central dataset.

5.4 Procedures and data collection

The empirical study consisted of three main steps: (i) preparatory work, (ii) a two-week asynchronous online deliberation exercise, and(iii) a follow-up questionnaire.

In the preparatory phase, students had four 2-hour seminars about: (a) Collective intelligence and online collaboration applications, (b) Argumentation theory, with a focus on IBIS and argument-based tools, (c) the Gold and Oil Markets, and (d) an instructional demo of Cohere or the Debate Dashboard. The students were also given a few reading materials and web links from specialised sources, such as economics-focused newspapers and magazines.. Additionally, as part of the preparatory phase, users went through warm-up exercises over the course of one week, in order to familiarise themselves with the tool by engaging in conversations around a different discussion topic (we did not use this output in our analysis).

At the start of the online deliberation, an initial map with a framing question and three mutually exclusive answers was presented to the subjects in both groups. In particular, the questions were:

What will be the trend of Gold (Oil) price in three months from now? The possible answers were: (i) The price will tend to increase (+10% or higher), (ii) The price will tend to decrease (-10% or lower), and (iii) The price will be stable (+/-10%). The two maps differed solely in the topic (Gold or Oil). When using the platform, participants were required to discuss and map contentious and/or competing points of view in argument maps, with alternative positions, and associated chains of pros and cons.

Due to various reasons, several students did not complete the activities. Eventually, we had 25 valid subjects in the treatment (Debate Dashboard) group and 39 in the control group. In principle, the different number of subjects involved in the two groups could have an impact on the users' performances; for instance, bigger groups could lead to more active and intense debates. While this might lead to greater risk of 'spaghetti' representation, it might also lead to better domain coverage. In fact, as detailed below, the data showed that the smaller group (treatment) was on average more active and produced more content. We also re-verified that the two groups were not significantly different in terms of demographics and academics after subjects in both groups had dropped out of the experiment ($t= 1.044$).

After the discussion, the participants completed a questionnaire (see Appendix A) composed of 25 multiple-choice questions (7-point Likert scale) articulated in four clusters aiming to assess the following constructs:*Mutual understanding, Perceived quality of collaboration, Accuracy of decision*and *Usability*.

5.5 Measurements

All constructs were measured through the post experiment questionnaire (Appendix A). With respect to*accuracy of decision*, participants simply reported their individual price forecast (i.e. the price will increase, be stable or decrease) and then directly compared the prediction with the actual value ex post. For the other three constructs, respondents indicated their level of agreement with a given statement by using an ordinal scale ranging from "Strongly disagree" to "Strongly agree" on a 7-point Likert scale.

- *Perceived Quality of Collaboration* was measured through 9 questions (Q1-Q9) identified by reviewing the literature on mediated collaboration (Daily-Jones et al., 1998; Sellen et al., 1992; Vandergriff, 2006).
- *Mutual Understanding* was measured through additional 9 questions constructed on the basis of a theliterature on grounding cost theory (Clark and Brennan, 1991) and studies of common ground building in mediated conversations (Monk and Watts, 2000; Convertino et al., 2007; 2008; 2009; Whittaker et al., 1998).
- *Usability* was assessed through 6 questions based on Davis' Technologies Acceptance Model (1989), and its subsequent modifications and extensions (Daily-Jones et al., 1998; Vassileva and Sun, 2007; Venkatesh, 2000; Venkatesh and Davis, 2000).

The questionnaire was validated to ensure content and construct validity (Appendix B).

6. Results

6.1 Comparison of the argument maps

In order to assess whether different types of feedback have an impact on a group's ability to explore the problem domain, we performed a content analysis comparing the argument map created by each group. For the sake of brevity, in the following we mainly focus on the outcomes of this analysis, while relevant methodological decisions are presented in Appendix C.

Two different rubrics, detailed in Appendix C, were developed to assess the quality of the two maps. The rubric *Quality of Argument Map*was designed to evaluate the structural quality of an argument map in terms of correct use of the argumentation formalism. It was based on the following criteria: (i) correct use of link type and direction (propositions), (ii) correct association of nodes to the appropriate IBIS category (examples), (iii) ability to support/attack arguments through evidence or counter-arguments.

The second rubric, *Quality of Domain Knowledge*, evaluated the quality of the content in terms of the group's ability to explore and provide adequate coverage of the problem space. This rubric is based on four criteria: (i) degree of exploration of the problem space; (ii) number of off-topic nodes; (iii) causality errors; and (iv) repetitions.

Quality of the argument map. Each argumentation move in each map was analysed as a *triple,*comprising two nodes and a meaningful connection between them (e.g. *challenges*). There were 415 triples generated in the Debate Dashboard group, and 380 triples in the control group. A comparison of the percentage of correct use of nodes type, links type, links direction, and appropriate use of evidence-based reasoning showed higher scores on four of the five indicators by the Debate Dashboard group (Table 2 and Figure 10). However, calculation of statistic significance on this measure did not confirm the hypothesis that the Debate Dashboard group would bestronger than the control group. Since the sample was large (more than 380 triples) we applied a Z-test for each of the criteria by comparing random pairs of argumentation triples in the two groups. Calculation of the Z-scores was never higher than the Z critical value (1.645 for 5% one tailed). We therefore cannot conclude that the Debate Dashboard feedback improved the quality of argumentation at the level of syntactic quality.

Table 2. Comparison between the two groups respect each argumentation quality criteria

	% of argumentation moves with appropriate use of link's types	% of argumentation moves with appropriate link direction	% of argumentation moves with appropriate category of posts	% of evidence-based arguments supporting arguments	% of evidence-based arguments challenging arguments
Control Group	90,53	91,84	73,68	68,95	21,84
Debate Dashboard	96,39	95,90	77,11	82,89	14,94

	% of argumentation moves with appropriate use of link's types	% of argumentation moves with appropriate link direction	% of argumentation moves with appropriate category of posts	% of evidence based arguemnts - supporting arguments	% of evidence based arguemnts - chellenging arguments
Control Group	90.5263158	91.8421053	73.6842105	68.9473684	21.8421053
Debate Dashboard	96.3855422	95.9036145	77.1084337	82.8915663	14.939759

Figure 10: Spider diagram showing the compared measures for each argumentation quality criteria.

It is noticeable that the percentage of counterarguments in both groups is much lower than other indicators, and that the control group is higher (both in absolute and percentage numbers) than in the Debate Dashboard group (Figure 10). While not significant statistically, this suggests a new hypothesis on the effect of social feedback on "disagreement", specifically, that there may be a negative correlation between the amount of social feedback provided to a group and the tendency of participants to express their disagreement publicly. Further research is needed to demonstrate whether higher social awareness leads to less disagreement, or agreater resistance to challenging others' ideas.

Quality of Domain Knowledge. Our analysis is based on the micro-economic model of market equilibrium in which several forces can influence the shift of demand and supply curves,thereby affecting the price trend of a commodity in the short term. These forces can be grouped into general factors such as the presence of substitute goods, production costs, demand demographics, etc., that do not depend on the specific type of good or service. For any market these factors follow the same causal relationships in influencing prices, for instance, an increase in the price of a substitute of product *A* will imply an increase in the demand for *A*. This approach allowed us to compare the content of the two maps even though they refer to two different commodities (Oil and Gold).

The quality of content was evaluated by considering the degree of coverage achieved by each group, measured in terms of the variables discussed and with what frequency. We also counted off-topic posts, causality errors, and the sheer replication of posts. A more detailed description of the evaluation criteria and their associated measures is reported in Appendix C (Table C.2).

In order to assess the quality of domain knowledge, a list of variables influencing market equilibrium was identified. Then, we mapped each post onto a list of keywords associated to each market force.

Two independent coders worked on a random sample of posts and their outcome was compared to check inter-rater reliability. By coding each post into a list of keywords, we obtained the distributions of posts across the list of keywords, as illustrated in Figure11.Through analysis of this distribution, it is possible to evaluate whether the discussion was balanced over a wide range of topics, or if instead participants were focused only on a subset of relevant terms.
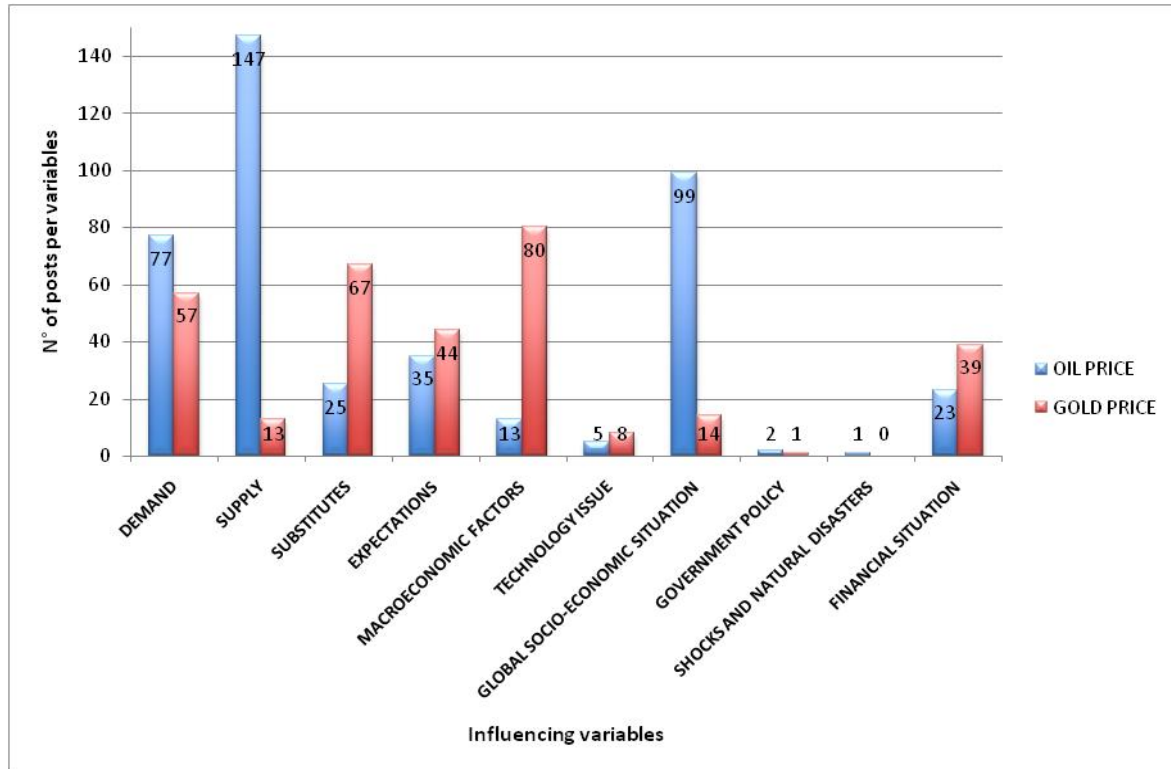
Figure 11. Distributions of created posts across relevant discussion topics
(blue = control, red = treatment)

Over two weeks, the students created about 600 posts. Figure 11shows that both groups were able to take into consideration several variables, ranging from economic, socio-political and financial aspects to technological and environmental issues. However, the participants in the control group focused on relatively few variables, while in the treatment group, users distributed their attention more evenly over the range of topics.

Table 3. Descriptive statistics about topics coverage by participant

Groups	Mean	Deviation standard
Treatment ($_T$)	5,1	1,92
Control ($_C$)	4	1,81
$t_{critical}(62)=1.671$; $t_{calculated}=2.509$; $\alpha=.05$		

Table 4. Off-topic, Repetition Rate and Wrong Causality values

Groups	Off-topic post (%)	Repetition Rate (%)	Causality error (%)
Treatment ($_T$)	2,97%	2,23%	0%
Control ($_C$)	4,79%	3,59%	0%

23

Table 4 shows that the treatment group outperformed the control group significantly also in terms of (less) off-topic and repeated posts, while causality errors were absent in both groups.

Accordingly, it is possible to claim that Debate Dashboard supported users in producing a better map in terms of content quality, while the structural quality, as measured through the rubrics adopted in this paper, was generally improved, but not significantly.

6.2 Users' activity

A straightforward measure of users' level of activity is the count of posts and connections created by each participant. During the experiment, both platforms were active 24 hours a day and we observed a high level of participation over time (Figures 12.1 and Figure 12.2). In particular, the users in the treatment group contributed significantly more than users in the control group on a day-to-day basis, except at the end of the experiment.
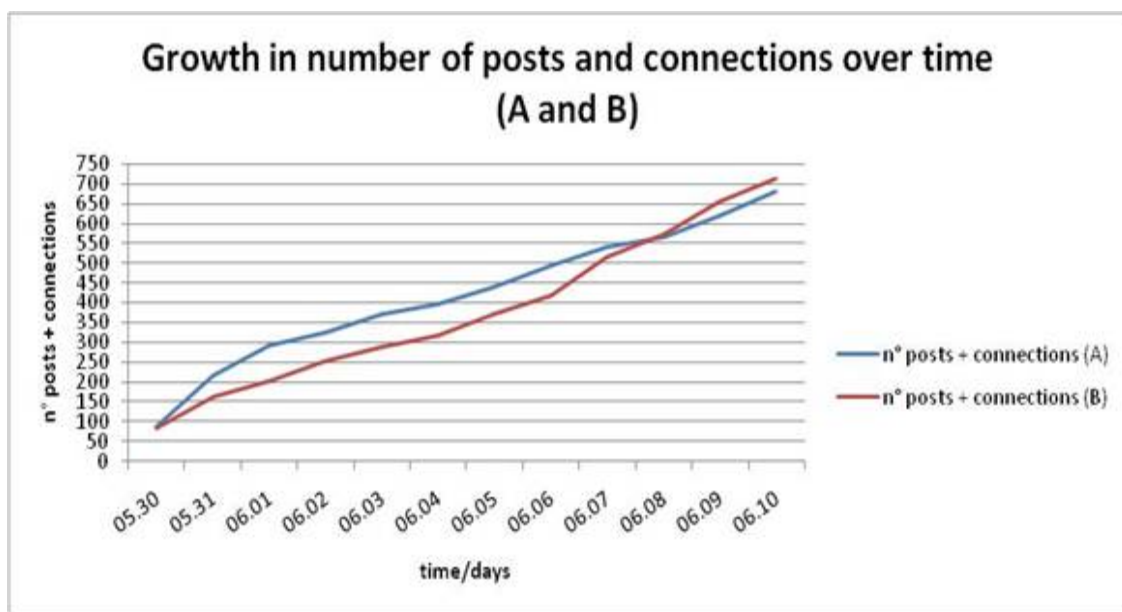
Figure 12.1: Growth in number of posts and connections over time in the treatment(A) and control group(B)

Table 5. Descriptive statistics on users' activity

	User's activity – Group A (treatment)	User's activity – Group B (control)
# Participants	25	39
# Posts	269	334
# Connections	412	380
Total	681	714
Mean Contributions/User	**27,24**	**18,31**
St. Deviation	22,98	16,28
$t_{critical}(62)=1.671$; $t_{calculated} =1.820$; $\alpha=.05$		

In two weeks the two groups posted 603 nodes and made 792 connections. Although students' participation may have been motivated by fact that the experiment was a course task for which they would have been evaluated by their professor, their informal face-to-face comments and results deriving from the analysis of the answers to the follow-up questionnaire showed that they found the

24

experiment interesting and appreciated the innovative features of the Debate Dashboard. Table 4 shows the number of posts and connections for each group. On average, Debate Dashboard users participated more, with around 27 posts per user compared with 18 contributions per user in the control group. The results of one-tail T-test confirm this difference is statistically significant (α= 0.05, Table 5).

Our data shows that the treatment group created also more connections (respectively, 412 vs 380). This can be offered as evidence that the Debate Dashboard users were more engaged.This increase in productivity may be due to a more efficient conversation facilitated by the availability of socially-salient meta-information.

The distribution of posts per user in both groups shows that level of activity was not homogeneously distributed among users, but that in both groups there was a small number of users that were considerably more active, as is typical in online communities (Wilkinson, 2008) (Fig. 13). While this pattern is common to both groups, the members of the Debate Dashboard group were in general more active, and there were more 'power users'in that group (24% of users with more than 40 contributions *vs.* 13% in the control group).
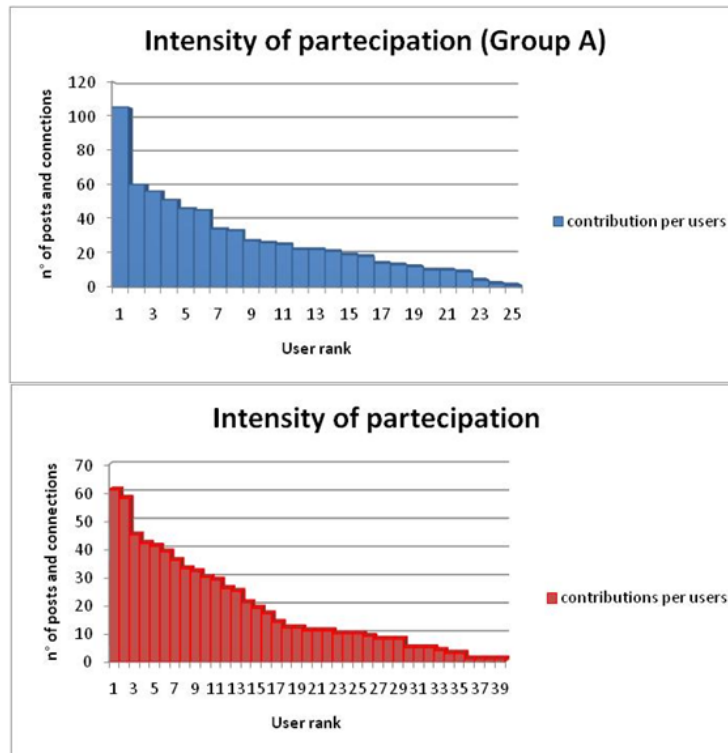
Figure 13. Number of contributions per user (Group A control, Group B experimental)

6.3 Collaboration process

In order to evaluate the impact of social and conversational feedback on collaborative performance, a range of statistical tests was performed to verify whether the two groups differed with respect to the following variables, as elicited by the questionnaire: *Mutual Understanding* (*MU*), *Quality of Collaboration* (*QofC*), *Quality of Decision* (*QofD*)and *Usability* (*Usab*). All the variables passed the normality test (Shapiro-Wilk, with p >0.05), so we used one tail t-tests for the following the set of hypotheses (H_0being the null hypothesis):

    Mutual Understanding: ***H$_0$:*** $MU_T = MU_C$

$$H_1: MU_T > MU_C$$

Quality of Collaboration:	$H_0: QofC_T = QofC_C$
	$H_2: QofC_T > QofC_C$
Quality of Decision:	$H_0: QofD_T = QofD_C$
	$H_3: QofD_T > QofD_C$
Usability:	$H_0: Usab_T = Usab_C$
	$H_4: Usab_T > Usab_C$

The results (Table 6) support most of thehypotheses, but at different levels of statistical significance. In particular, it was possible to reject the null hypothesis (H_0) at $\alpha = .05$ for *Mutual Understanding* and perceived *Quality of Collaboration*. Instead, for *Quality of Decision* evidence is weaker and we can reject H_0 only at $\alpha = .1$. Finally, with regard to *Usability* we cannot reject H_0, and the usability ratingsarenot significantlydifferent.

Table 6. Questionnaire ratings of collaboration process

Latent variables	Group A (control)		Group B (experimental)	
	Mean	S.D.	Mean	S.D.
MU	34.03	5.17	36.68*	5.21
QofC	41.03	5.63	43.68*	4.89
QofD	44.72	5.49	47.96**	5.06
Usability	30.27	6.30	31.68	3.87
*p=.05; **p=.1				

In order to measure the magnitude and direction of the difference between theconditions, we computed the effect size (ES) through Cohen's *d* (Cohen, 1988). The results reported in Table 7 are evidence for a moderate effect on *Mutual Understanding* and *Quality of Perceived Collaboration*, while we see a small effect for *Usability* and *Quality of Decision*.

Table 7. Effect size tests

Effect size	
Mutual Understanding	0.51
Quality of Collaboration	0.50
Quality of Decision	0.40
Usability	0.26

7. Discussion

The aim of this work is to contribute to the design of better collaborative tools to support geographically distributed groups in making sense of complex problems, through sharing and debating ideas. In the following we identify and discuss several implications deriving from this study.

Socially-augmented group deliberation. Our study offers evidence that the socially augmented version of the CCSAV tool used in our experiment was able to improve *Mutual Understanding*, *Quality of Collaboration* and, in a more limited manner, *Quality of Decision*. The users of the treatment group were more engaged, productive, and achieved better coverage of the problem space, although we observed no significant improvement in the use of the argumentation formalism. These findings suggest that the design of artefact-centred collaborative tools should take into account the delivery of social and conversational meta-information, whereas, to date,most work on CSAV has emphasised the importance of knowledge representation and the quest of adequate argument ontologies (Scheuer et al., 2010), while neglecting the issue of how to make interaction more socially meaningful to users.

We can offer only limited evidence that social augmentation helps to produce better maps; in particular, we observed a significant improvement in the users' ability to explore the problem space, while the structural quality of the map did not increase significantly. More in-depth and sophisticated data analysis should be performed to understand the relationship between social augmentation, mediation capability and the production of better knowledge artefacts. However, our findings show that social augmentation did not come at the expense of artefact quality. Improvements, though limited, to the quality of content and individual decisions obtained in absence of amelioration of the structural quality of the artefact, may suggest that the importance of the formalism and of the capability of participants to use it proficiently to build well-formed knowledge representations,need not be the exclusive priority for designers.

Social and conversation meta-information may help users to overcome, at least in part, their relative lack of argument mapping skills. In other words, having artefact-centred discussions that are more socially rich and meaningful can help make CCSAV platforms productive even when used by unskilled or untrained participants. A possible explanation is that socially connected participants are able to develop a sense of joint understanding and feel less frustrated by their inability to use the tool properly. This of course would not prevent trained and skilled users from using the tool evenmore effectively. It would be interesting to replicate our study with users characterised by different levels of argument mapping skills, and test for the effects of such proficiency.

Artefact-centred online interaction. Our work suggests a design approach to deliver social and conversational meta-information in artefact-centred collaborative tools. While previous studies have generally shown the importance of social translucence in the design of collaborative tools (e.g. Erikson and Kellogg, 2002), to our knowledge there have been no systematic attempts to develop design solutions that link meta-information to the collective production of a knowledge artefact. The meta-informational feedback loops we have introduced are created by the system in real time and are a by-product of the interaction that happens through the deliberation map. Feedback delivery is obtained through a set of ancillary representations available in the background, while the knowledge object that participants are supposed to create during the discussion stays clearly on the forefront.

In the design of the Debate Dashboard we have carefully avoided the use of multiple, parallel discourse representations. For instance, onepossible choice would have been to provide users with both a conversational space, in which the communication would be kept fluid and informal, and an object space, in which the conversation should solidify into a knowledge object (an argument map in this case). Such a workflow would be similar to the creation of Wikipedia articles, in which users switch between the collaborative output (the encyclopaedia definition) and the collaborative process (the discussion page). We decided to not follow this strategy for several reasons.

First, we suspect that the gap between an online argument map and an online conversation in terms of their ability to mediate interaction is quite large, and definitely much wider than the one occurring between a wiki collaborative text and its behind-the-scene conversation. If our intuition is correct, chances are high that in the case of an artefact-centred tool, users will eventually focus their attention and effort on the interaction space that is more convenient to them, which would presumably be the online conversation. Of course, this conjecture would require additional empirical evidence to be validated.

Second, our research objective was to stress our CCSAV platform in terms of its ability to mediate interaction and communication without compromising its main feature.First and foremost, CCSAV platforms are tools in which the knowledge artefact should be the main focus and locus of interaction. In other words, we were interested in analysing to what extent users were able to interact and talk *through the map*, while it was of little interest to us to observe users talking *about a map*.

8. Conclusions

8.1 Contribution

In this work we have presented a socially augmented CCSAV platform, the Debate Dashboard, designed through the integration of an existing argument mapping tool (Cohere) and a set of additional representations that deliver social and conversational meta-information. This paper aims to contribute to the debate around the use and design of web-based argumentation technologies to support distributed knowledge tasks such as group deliberation.

We have criticised existing argument-based tools for their limited capability of mediating interaction among users involved in online conversations. By drawing upon previous studies in conversational analysis (Clark and Brennan, 1991), online interaction in computer-supported cooperative work (Erikson and Kellogg , 2000; McDonald et al., 2009), and communities of practice (Wenger, 1998), we have juxtaposed argumentation and conversation along a continuum, and have argued that artefact-centred technologies such as CCSAV platforms introduce high grounding costs because they neglect the role of a range of conversational and social feedback that are needed by participants to build mutual understanding through conversational exchanges.

Our empirical findings show that a critical limitation of CCSAV technologies lie in their inability to mediate interaction, due to the lack of access to socially salient information that is available in face-to-face conversations. To the best of our knowledge, no previous studies have integrated argument-mapping tools with social meta-information. Our empirical findings show that the development of common ground could be a crucial element for successful collaborative processes using CCSAV platforms.

8.2 Limitations and future work

One limitation of this research is that experimental subjects were academic students working in a semi-controlled field test, which limits the external validity of our study. On the other hand, it was too costly and risky to design a field test involving professional users of CCSAV tools at this stage of development of the theory and our prototype. To address this limitation, we hope to be able to replicate the study with a real-world online community involving experts and practitioners. External validity may be also affected negatively by the fact that the results have been obtained in relation to a specific platform. While it is costly and difficult to carry out parallel empirical studies that may disentangle the effect of diverse tools from the theory under test, we plan to replicate our work in

different settings characterised by the use of other tools with the objective of developing multiple case studies.

A limitation of many research prototypes is the lack of polish in the user experience. It is reasonable to assume that this has affected to some extent the usefulness of these widgets, as well as the empirical results. Unfortunately, the implementation of a more refined prototype was not affordable at this stage of the research because a full implementation would have required deep and time-consuming changes to the underlying pre-existing deliberation tool (Cohere). However, despite being an early prototype,many hypotheses were confirmed, so we may speculate that a more refined and usable design would yield results thatimprove upon our findings.

Finally, although our results show that social and conversational feedback have a positive impact on users' performance, more in depth analysis should be carried out to better understand whether and how each type of feedback impacts mutual understanding and the collective deliberation processes.

The next step in our work will be the implementation of a more complete and refined prototype that will deliver richer feedback than the one we have currently implemented. Since the Debate Dashboard was implemented on the top of an existing asynchronous tool, many needed features that are available in synchronous platforms were missing and, in particular, we were forced to leave out several important types of interaction feedback (co-temporality, simultaneity and users' mobility). In the next release we aim to cover all of the constraints identified by Clark and Brennan (Table 1, Fig. 1).

Finally, we plan to carry out additional research into the nature and the structure of the cause-effect relationship between feedback use and deliberation performance. For instance, it would be interesting to analyse users' behaviour over time to observe how they move between content and meta-information, e.g. how much time they spend on different types of feedback, and whether there are patterns in the way subjects consult and utilise different sources of feedback. A better understanding of how different forms of feedback are used in practice and how their use affects performance would make our theoretical framework stronger and at the same time produce more concrete design implications.

Acknowledgements

References

Andriessen, J., Baker, M., Suthers, D. 2003. Arguing to learn: Confronting cognitions in computer-supported collaborative learning environments. Dordrecht, Netherlands: Kluwer Academic Publishers. Computer-supported collaborative learning book series, vol 1.

Bell, P. 1997. Using argument representations to make thinking visible for individuals and groups. In Hall R., Miyake N., and Enyedy N. (Eds.), In CSCL'97 Proceedings of the 2nd International conference on Computer Supported Collaborative Learning, 10-20.

Borgatti, Stephen P., Everett, M. G. 2005.A Graph-Theoretic Perspective on Centrality. Social Networks, 28, 466–484

Brna, P., Cox, R., Good, J. 2001. Learning to think and communicate with diagrams: 14 questions to consider. Artificial Intelligence Review, 15(1-2), 115-134.

Bryant, F.B. 2000. Assessing the Validity of Measurement, in: Grimm, L.G., and Yarnold, P.R. (Eds.) Reading and Understanding More Multivariate Statistics, Washington DC: American Psychological Association, 99-146.

Bruffee, K.A. 1993. Collaborative Learning: Higher education, interdependence, and the authority of knowledge. John Hopkins University Press, Baltimore.

Buckingham Shum, S. 2003. The roots of computer supported argument visualization. In Kirschner, P.A., Buckingham Shum, S., and Carr, C.S. (Eds.) Visualizing Argumentation: Software Tools for Collaborative and Educational Sense-Making, Springer-Verlag, 3-24.

Buckingham Shum, S. 2008. Cohere: Towards Web 2.0 Argumentation. Proceedings of the 2nd International Conference on Computational Models of Argument. 28-30 May 2008, Toulouse, France: IOS Press, 97-108.

Buckingham Shum S and Okada A. (2008) Knowledge Cartography for Controversies: The Iraq Debate. In: Okada A, Buckingham Shum S and Sherborne T (eds) *Knowledge Cartography: Software Tools and Mapping Techniques.* London: Springer.

Buckingham Shum, S., Hammond, N. 1994. Argumentation-based design rationale: What use at what cost? International Journal of Human-Computer Studies, 40(4), 603-652.

Carr, C.S. 2003. Using computer supported argument visualization to teach legal argumentation. In Kirschner, P.A., Buckingham Shum, S., and Carr, C.S. (Eds.) Visualizing Argumentation – Software Tools for Collaboration and Educational Sense-making, Springer-Verlag, London, 75-96.

Chin W.W. 1998. The partial least squares approach to structural equation modelling. Modern methods for business research, 295-336.

Cho, K.L. Jonassen, D.H. 2002. The effects of argumentation scaffolds on argumentation and problem solving. Educational Technology Research and Development, 50(3), 5-22.

Clark, H.H., Brennan, S.E. 1991. Grounding in communication, in Resnick, L.B.L.,Teasley J.M., and Stephanie, D. (Eds.), Perspectives on socially shared cognition, Washington (DC): American Psycho Kieslerlogical Association, pp. 127-149.

Cohen, J. 1988. Statistical power analysis for the behavioral sciences. Lawrence Erlbaum.

Conklin, J. and Begeman, M.L. 1988.gIBIS: A Hypertext Tool for Exploratory Policy Discussion. ACM Transactions on Office Information Systems, 4(6), 303-331.

Conklin, J. 2003. Dialog Mapping: Reflections on an Industrial Strength Case Study, in Kirschner, P.A., Buckingham Shum, S., and Carr, C.S. (Eds.), Visualizing Argumentation – Software Tools for Collaboration and Educational Sense-making, Springer-Verlag, pp. 117-131.

Conklin, J. 2006. Dialogue Mapping: Building Shared Understanding of Wicked Problems. Chichester (UK): Wiley.

Convertino, G., Hong, L., Nelson, L., Pirolli, P., Chi, E.H. 2009. Activity Awareness and Social Sensemaking 2.0: Design of a Task Force Workspace, in Lecture Notes in Computer Science, Foundations of Augmented Cognition. Neuroergonomics and Operational Neuroscience, 2009, pp. 128-137.

Convertino, G., Mentis, H.M., Rosson, M.B., Carroll, J.M., Slavkovic, A., Ganoe, C.H. 2008. Articulating common ground in cooperative work: content and process, in CHI'08, Proceeding of the twenty-sixth annual SIGCHI conference on Human factors in computing systems, ACM New York, NY, USA.

Convertino, G., Mentis, H.M., Ting, A.Y.W.,Rosson, M.B., Carroll, J.M 2007. How does common ground increase?.Proceedings of the 2007International ACM conference on Supporting group work, 225-228.

Convertino, G., Neale, D.C., Hobby, L., Carroll, J.M., Rosson M.B. 2004. A laboratory method for studying activity awareness, In NordCHI'04 Proceedings of the third Nordic conference on Human-computer interaction, ACM New York, NY, USA.

Cronbach, L.J. 1951. Coefficient alpha and the internal structure of tests.Psychometrika, 16(3), 297-334.

Daily-Jones, O., Monk, A., Watts, L. 1998. Some advantages of video conferencing over high quality audio conferencing: Fluency and awareness of attentional focus. International Journal of Human-Computer Studies, 45, 21-58.

Davis, F.D. 1989. Perceived usefulness, perceived ease of use and user acceptance of information technology.MIS Quarterly, 13(3), 319-339.

De Liddo, A., Buckingham Shum, S. 2010. Cohere: A prototype for contested collective intelligence. Workshop on Collective Intelligence In Organizations - Toward a Research Agenda, February 6-10, 2010, ACM Computer Supported Cooperative Work (CSCW 2010)Savannah, Georgia, USA.

De Liddo, A., Buckingham Shum, S., Quinto I., Bachler, M. and Cannavacciuolo, L. 2011.Discourse-Centric Learning Analytics. Proc. 1st International Conference on Learning Analytics & Knowledge. Banff, CA: ACM: New York, 22-33. DOI: http://dx.doi.org/10.1145/2090116.2090120 Open Access Eprint: http://oro.open.ac.uk/25829

Dede, C. 1996. The Evolution of Distance Education: Emerging technologies and distributed learning. American Journal of Distance Education, 10(2), 4-36.

Englemann, T., and Hesse, F.W. 2010.How digital concept maps about the collaborators' knowledge and information influence computer-supported collaborative problem solving.Computer-Supported Collaborative learning, 5, 299-319.

Erickson, T., Kellogg, W.A. 2000. Social translucence: an approach to designing systems that support social processes. ACM Transaction on Computer-Human Interaction, 7(1), 59-83.

Erickson, T., Kellogg, W.A. 2002. Knowledge Communities: Online Environments for Supporting Knowledge management and its Social Context, in Ackerman M.A., Pipek, V., and Wolf, W. (Eds.) Beyond knowledge management: Sharing expertise, MIT Press, Cambridge, MA.

Fornell, C., Larcker, D.F. 1981. Evaluating structural equation models with unobservable variables and measurement error. Journal of marketing research, 39-50.

Gefen, D., Straub, D. 2005. A practical guide to factorial validity using PLS-graph: Tutorial and annotated example. Communications of the Association for Information Systems, 16(1), 91-109.

Gürkan, A., Iandoli, L., Klein, M., Zollo, G. 2010.Mediating debate through on-line large-scale argumentation: Evidence from the field. Information Sciences, 180, 3686-3702.

Hair, D.C. 1991. Legalese: A legal argumentation tool. SIGCHI Bulletin, 23(1), 71-74.

Hewitt, J. 2001. Beyond threaded discourse. International Journal of Educational Telecommunications, 7(3), 207-221.

Iandoli, L, Klein, M. and Zollo, G. 2009. Enabling on-line deliberation and collective decision-making through large-scale argumentation: a new approach to the design of an Internet-based mass collaboration platform. International Journal of Decision Support System Technology 1, 69-91.

Introne, J.E. 2009. Supporting group decisions by mediating deliberation to improve information pooling. In Proceedings of the ACM 2009 International Conference on Supporting group work, 189-198.

Karacapilidis, N., Tzagarakis, M., Karousos, N., Gkotsis, G., Kallistros, V., Christodoulou, S., Mettouris, C., Nousia, D. 2009. Tackling cognitively-complex collaboration with CoPe_it!.International Journal of Web-Based Learning and Teaching Technologies, 4(3), 22-38.

Kirschner, P.A., Buckingham Shum, S., Carr, C.S. (Eds.) Visualizing Argumentation: Software Tools for Collaborative and Educational Sense-Making, Springer-Verlag, 3-24

Klein M. (2012).Enabling Large-Scale Deliberation Using Attention-Mediation Metrics. Computer Supported Cooperative Work, 21, 449-473.

Kuhn, D. 1991.The skills of argument. New York: Cambridge University Press.

Loui, R.P., Norman, J., Altepeter, J., Pinkard, D., Craven, D., Linsday, J., Foltz, M., (1997). Progress on Room 5: a testbed for public interactive semi-formal legal argumentation. In Proceedings of the 6th international conference on Artificial intelligence and law, 207-214.

Mercier, H., Sperber, D.2011.Why do humans reason? Arguments for an argumentative theory. Behavioral and Brain Sciences, 34, 57–111.

McCarthy, J.C., Miles, V.C., Monk, A.F. 1991. An experimental study of common ground in text-based communication. In Proceedings of the SIGCHI conference on Human factors in computing systems: Reaching through technology, 209-215.

McDonald, D.W., Beschastnikh, I., Kriplean, T., Borning, A., Zachry, M. 2009, System Design for Social Translucence in Socially Mediating Technologies. Socially Mediating Technologies

Workshop at the ACM 2009 SIGCHI Conference on Human Factors in Computing Systems (CHI'09).

Monk, A., Watts, L. 2000. Peripheral Participation in Video-Mediated Communication.International Journal of Human-Computer Studies, 52 (5), 933-958.

Novak, J.D. and Gowin, D.B. 1984. Learning how to learn. New York: Cambridge University Press.

Okada, A., Buckingham Shum, S. 2008. Evidence-Based Dialogue Maps as a research tool to evaluate the quality of school pupils' scientific argumentation. International Journal of Research and Method in Education, 31(3), pp. 291–315.

Okada, A., Buckingham Shum, S., Sherborne, T. 2008. Knowledge Cartography: software tools and mapping techniques. Springer-Verlag, London.

Pardo A and Kloos CD. (2011) Stepping out of the box: towards analytics outside the learning management system. *Proceedings of the 1st International Conference on Learning Analytics and Knowledge.* Banff, Alberta, Canada: ACM, 163-167.

Rahwan I. and Simari G.R. (2009) (Eds.) *Argumentation in Artificial Intelligence.* Springer.

Reed C. and Rowe G. (2004) Araucaria: Software for Argument Analysis, Diagramming and Representation. *International Journal on Artificial Intelligence Tools* 13: 983.

Rittel, H.W.J., Webber, M.M.1973. Dilemmas in a general theory of planning.Policy sciences, 4(2), 155-169.

Roschelle, J. 1996. Designing for cognitive communication: Epistemic fidelity or mediating collaborating inquiry. In Day, D.L. and Kovacs, D.K. (Eds.) Computers, Communication and Mental Models, Taylor & Francis, London, 13–25.

Royal, M.A., Rossi, R.J. 1996. Individual-level correlates of sense of community: Findings from workplace and school. Journal of Community Psychology, 24(4), 395-416.

Scheuer, O., Loll, F., Pinkwart, N., McLaren, B.M. 2010. Computer-supported argumentation: A review of the state of the art. Journal of Computer-Supported Collaborative Learning, 5(1), 43-102.

Shim, J.P.,Warkentin, M., Courtney, J.F., Power, D.J., Sharda, R., Carlsson, C. 2002. Past, present, and future of decision support technology. Decision Support Systems, 33(2), 111-126.

Sellen, A.J. 1992. Speech Patterns in Video-Mediated Conversations. In CHI '92: Human Factors in Computing Systems, 49-59.

Selvin, A.M., Buckingham Shum, S.J. and Aakhus, M. 2012The practice level in participatory design rationale: studying practitioner moves and choices. In: Carroll, John M. ed. Creativity and Rationale: Enhancing Human Experience by Design. London: Springer, pp. 287–325.(Reprinted from: Human Technology: An Interdisciplinary Journal of Humans in ICT Environments, 6(1) pp. 71–105.)

Sierhuis, M. and Buckingham Shum, S. 2008.Human-agent knowledge cartography for e-science: NASA field trials at the Mars Desert Research Station. In: Okada, Alexandra; Buckingham Shum, Simon and Sherborne, Tony (eds.)Knowledge Cartography: Software Tools and Mapping

Techniques. Advanced Information and Knowledge Processing Series. London: Springer, pp. 287–306

Suthers, D. Hundhausen, C.D., Girardeau L. E. 2003. Comparing the roles of representations in face-to-face and online computer supported collaborative learning. Computers & Education 41(4): 335-351.

Suthers, D., Weiner, A. 1995. Groupware for developing critical discussion skills. Paper presented at CSCL'95: Computer Supported Collaborative Learning, Bloomington, IN.

Suthers, D.D., Vatrapu, R.K., Medina, R., Joseph, S.R.J, Dwyer, N. 2008. Beyond threaded discussion: Representational guidance in asynchronous collaborative learning environments. Computers & Education 50(4): 1103-1127.

Toulmin, S. 1958. The use of Arguments, Cambridge University Press, Cambridge (MA).

Twardy, C. 2004. Argument Maps Improve Critical Thinking. Teaching philosophy, 2, 95-116.

van den Braak, S.W., van Oostendorp, H., Prakken, H., Vreeswijk, G.A.W. 2006. A critical review of argument visualization tools: do users become better reasoners?, in Working Notes of the 6th Workshop on Computational Models of Natural Argument (CMNA2006), pp. 67-75.

van Gelder, T. 2003. Enhancing deliberation through computer supported argument visualization, in Kirschner, P.A., Buckingham Shum, S. and Carr, C.S. (Eds.) Visualizing Argumentation: Software Tools for Collaborative and Educational Sense-Making, Springer, London, UK, pp. 97-115.

van Gelder, T. 2007. The rationale for Rationale™. Law, Probability and Risk, 6(1-4), 23-42.

Vandergriff, I.2006. Negotiating common ground in computer-mediated versus face-to-face discussions. Language Learning & Technology, 110-138.

Vassileva, J., Sun, L. 2007. An improved design and a case study of a social visualization encouraging. Groupware: Design, Implementation, and Use, 72-86.

Venkatesh, V. 2000. Determinants of perceived ease of use: Integrating control, intrinsic motivation, and emotion into the technology acceptance model. Information Systems Research, 342-365.

Venkatesh, V., Davis, F.D. 2000. A theoretical extension of the technology acceptance model: Four longitudinal field studies. Management science, 186-204.

Walton, D. 2008. Informal logic: a pragmatic approach. Second Edition, Lavoisier Libraire.

Walton D, Reed C and Macagno F. (2008) Argumentation Schemes, Cambridge: Cambridge University Press.

Wellman, B. 1999. Networks in the global village: Life in contemporary communities. Westview Press.

Wenger, E. 1998.Communities of practice: Learning, meaning, and identity. Cambridge University Press.

Whittaker, S., Terveen, L., Hill, W., Cherny, L. 1998. The dynamics of mass interaction. In Proceedings of the 1998 ACM conference on Computer supported cooperative work, 257-264.

Wikes-Gibbs, D., Clark, H.H. 1992. Coordinating beliefs in conversation. Journal of Memory and Language. 31(2), pp. 183-194.

Wigmore, J.H. 1913. The principles of judicial proof : as given by logic, psychology, and general experience, and illustrated in judicial trials. Boston: Little, Brown.

Wilkinson, D.M. 2008. Strong regularities in online peer production.In EC '08 Proceedings of the 9th ACM conference on Electronic commerce, New York, NY.

Appendix A – Questionnaire items, constructs, and main sources

(English translation from Italian)

	Questionnaire items	Constructs	Sources
Q1	The interaction level developed during Cohere-mediated conversation was satisfying related to the aim of the online discussion	Quality of online Collaboration	Sellen *et al.,* 1992 Vandergriff, 2006; Daily-Jones *et al.,* 1998
Q2	I found the online discussion interesting and engaging		
Q3	Collaboration was effective to solve the assigned problem		
Q4	I found it difficult to keep track of the conversation		
Q5	The argument map was helpful in facilitating knowledge sharing among team members		
Q6	I shared my own knowledge about the task with my teammates		
Q7	I found that my teammates have shared their own knowledge about the task		
Q8	The online community developed a good amount of work		
Q9	The online community made a good job		
Q10	I think that, at the end of online debate, the group has a common position about the discussion topic	Quality of Decision	Convertino *et al.,* 2007 Vandergriff, 2006.
Q11	What was your decision at the end of online debate?		
Q12	According to you, if the group had to make a collective decision (e.g. voting), what would be its decision after the discussion?		
Q13	In general, I have not had problems to understand the meaning of other members' posts	Mutual Understanding	Monk and Watts, 2000 Whittaker *et al.,* 1998 Convertino *et al.,* 2004; 2007; 2008; 2009 McCarthy *et al.,* 1991
Q14	In general, I think that the other members have understood my contributions without difficulty		
Q15	I could easily sum up the key arguments and the key positions developed during the online discussion		
Q16	I could easily identify who has done what		
Q17	I could easily say who is online on Cohere		
Q18	My teammates and I developed better understanding about others' ideas and opinions about the topic over the two weeks		
Q19	My teammates and I developed shared understanding about the task over the time		
Q20	I found online conversation was often redundant		
Q21	I found there are many irrelevant posts respect to the assigned task		
Q22	Interaction with the system does not require me a lot of mental effort	Usability	Venkatesh and Davis, 2000; Vassileva and Sun, 2007; Convertino *et al.,* 2007; Venkatesh, 2000
Q23	I find the system easy to use		
Q24	I enjoyed collaborating with my teammates using Cohere		
Q25	I would be interesting to re-use Cohere in similar applications/works		
Q26	It was easy to communicate effectively with the other team members by using the tools available and its widgets		
Q27	Cohere makes collaboration among online users more fluid and quick		

Appendix B – Content and Construct validity

Content validity

Most of the survey items have been adapted from existing, validated scales used in the literature on mediated conversation (Convertino et al., 2008; Convertino et al., 2009; Monk and Watts, 2000; Whittaker et al., 1998; McCarthy et al., 1991), online collaboration (Daily-Jones et al., 1998; Sellen et al., 1992; Vandergriff, 2006) and the Technology Acceptance Model (Davis, 1989; Venkatesh, 2000; Venkatesh and Davis, 2000). However, additional tests were made during the field study to ensure content validity with a small group of fifteen respondents. The respondents' feedback was used to improve the face-validity of the questionnaire, revise wording, and eliminate ambiguities.

Construct validity

Empirical tests were used to examine the measurement properties of the indicators and to establish *reliability, convergent analysis* and *discriminant validity.*

With the regard to *reliability,* we computed Cronbach's α(1951). After eliminating some questions we obtained the values reported in table B1. As showed in the table, the Cronbach's α values exceed the minimum recommended value (0.7).

Table B.1 Cronbach's alpha

Constructs	Cronbach's α
Mutual Understanding	0.763
Quality of Collaboration	0.744
Quality of Decision	0.752
Usability	0.845

Convergent validity is the degree to which multiple measures of the same construct demonstrate agreement or convergence (Bryant, 2000, p. 113). Convergent validity is attained when multiple measures of an item represent the same underlying construct. In this study, convergent validity was examined by computing the index of Average Variance Extracted (AVE), as proposed by Fornell and Larker (1981). AVE measures the amount of variance captured by the indicators in relation to the amount of variance due to measurement error (Fornell and Larcker, 1981). In other words, AVE is a measure of the error-free variance of a set of items. AVE should be greater than 0.5 to provide support for convergent validity (Chin, 1998; Fornell and Larcker, 1981), which is the case in this study (Table B.2).

Table B.2 Average Variance Extracted values

Constructs	Average Value Extracted
Mutual Understanding	0.953
Quality of Collaboration	0.898
Quality of Decision	0.912
Usability	0.940

Discriminant validity measures the extent to which measurements of different concepts are distinct (Bryant, 2000). Discriminant validity was assessed using the method prescribed by Gefen and Straub (2005). The procedure to assess discriminant validity is again AVE analysis, performed by comparing the square root of the AVE with the inter-constructs correlation. The square root of the AVE has to be larger than the correlations with the other constructs (Fornell and Larcker, 1981). Unfortunately, there are no definitive guidelines to indicate how much larger the square root of the AVE should be. In our case, the square root of the average variance extracted is much larger than the correlations of the construct with the all of the other constructs (Table B.3), which is considered an acceptable criterion to pass the discriminant validity test.

Table B.3. Discriminant validity

CONSTRUCT	Sqrt AVE	SIC
Mutual Understanding	0,909	0,564; 0,234; 0,805
Quality of Collaboration	0,807	0,564; 0,234; 0,624
Quality of Decision	0,832	0,234; 0,234; 0,199
Usability	0,883	0,624; 0,805; 0,199

Appendix C - Rubrics to compare Argument Maps

Rubrics are a scoring method based on the description of students' expected performance for a given assignment. In general, rubrics are presented as detailed descriptions of all the criteria and measures that an instructor will use to evaluate the performance. Criteria are defined either heuristically or by relying on theories of what a "gold standard" performance looks like for a specific assignment.

Whereas in common teaching practice rubrics are used to provide feedback to students on their performance, in this paper we use rubrics to assess the quality of the argument maps produced by the two groups of students. In a similar fashion, rubrics have been widely used in the literature as a method to assess Concept Maps (e.g., Novak and Gowin, 1984) based on the components and structure of the map. Since Argument maps have different features and standards than concept maps, we adapted and extended Novak and Gowin's rubrics to the case of argument maps.

Each of the rubrics we use is comprised of three main elements: *criteria, descriptors* and *performance levels*. Criteria represent the dimensions that will be measured and this includes a definition and some examples to clarify the interpretation of that criteria. Performance levels usually take the form of adjectives and describe the degree of performance.

Rubric 1 – Quality of argument map

This rubric is designed to measuring the quality of an argument map in terms of proper use of argument formalism and ontology. We applied this rubric to assess the work of each student, taking into account all the argumentation triples (two ideas and their semantic connection) that this student was involved in creating. Each semantic triple has been assessed following the criteria reported in table C.1.

Table C.1:*Quality of Argumentation* rubric consisting of 5 criteria adapted from (Novak and Gowin 1984)

Rubrics	Criteria	Descriptors	Measures and Ratings
Quality of Argumentation	*Propositions – Link label*	This criterion verifies the appropriate use of propositions (connections between nodes in an argument map). Connecting labels are used to explain the rhetorical or semantic relationship between two ideas.	% of links with appropriate use of link types. Evaluation scale: Excellent (75-100%); Good (50-75%); Poor (25-50%); very poor (0-25%).
	Propositions - Link Direction	This criterion checks if there was an appropriate use of connection between nodes in the argument map, from the point of view of the direction of the connection between nodes.	% of links with appropriate link direction. Evaluation scale: Excellent (75-100%); Good (50-75%); Poor (25-50%); very poor (0-25%).
	Examples (Node Types)	This criterion evaluates whether nodes in the map are classified in the appropriate category (Questions/Answer/Pro/Con).	% of nodes within the appropriate category of posts. Evaluation scale: Excellent (75-100%); Good (50-75%); Poor (25-50%); very poor (0-25%).

	Evidence-based reasoning	This criterion evaluates the use of evidence to back up a contribution (*Pro*).	% of evidence based arguments. Evaluation scale: Excellent (75-100%); Good (50-75%); Poor (25-50%); verypoor(0-25%).
	Counter Arguments	This criterion evaluates the use of counter-arguments in the map to challenge a contribution (*Con*).	% of counter-arguments. Evaluation: Excellent (75-100%); Good (50-75%); Poor (25-50%); very poor (0-25%).

Rubric 2 – Quality of domain knowledge

With this rubric we assess a students' ability to explore the domain knowledge that is required to discuss and predict the price of the two commodities. Each contributed node was evaluated on the basis of the criteria and standards described in the table below.

Table C.2:*Quality of Domain Knowledge* rubric consisting of 4 criteria

Rubrics	Criteria	Descriptors	Measures & Ratings
Quality of Domain Knowledge	*Domain coverage*	This criterion evaluates the use of variables from the reference causal model (market equilibrium)	# nodes that contain variables of the causal model
	Off-topic nodes	This criterion refers to the inclusion of irrelevant topics	# of nodes that deal with irrelevant topics
	Causal errors	This criterion evaluates the correct use of the "sign" of the causal effects in the reference model	# of "wrong causality" mistakes
	Repetition rate	This criterion measures redundancy as sheer replication of a same content in the map	# of repeated nodes

In order to assess the domain coverage, we used as a reference the model of market equilibrium, through the identification of market forces influencing equilibrium:
- *Demand* (variation on the demand of crude oil/gold)
- *Supply* (variation in the supply of crude oil/gold)
- *Substitutes* (e.g. renewable energy, natural gas)
- *Expectations of rising/decreasing price* (e.g. expectations of rising price, forecasts by famous economists etc.)
- *Macroeconomic factors* (i.e. interest rate, inflation, gold reserve by Central banks, monetary policy)
- *Technology issue* (i.e. technological innovation, costs)
- *Global socio-economic situation* (i.e. economic crisis, politics)
- *Government policy* (i.e. incentives)
- *Shocks and Natural Disaster* (war, earthquakes, hurricanes, etc.)
- *Financial situation* (i.e. financial crisis, financial market quotations, etc.).

We coded each post into a list of keywords through labels related to the above categories.

Author biographies

Luca Iandoli is Professor of Engineering Management at the Univ. of Naples Federico II (Italy) and at the Stevens Institute of Technology (USA), and a former Fulbright Fellow at the Center for Collective Intelligence, MIT. His research activities focus in the following areas: Knowledge Management, Organizational Learning, Collaborative technologies in organizations and networks of small firms. He has published many papers on the analysis of collaborative dynamics in firms' networks through computational methodologies (agent-based modeling, fuzzy logic, social network analysis). He is President for 2011/2012 of the European Council for Small Business and entrepreneurship (ECSB), the largest European academic association on entrepreneurship and small business management research.

Ivana Quinto received her PhD in Science and Technology Management from the University of Naples Federico II in 2012. Her research focuses on how web-based collaborative technologies can offer better support for knowledge sharing and collective deliberation processes. Prior to finishing her degree, she was a visiting student at the Knowledge Media Institute of the Open University (Milton Keynes, UK) a 70-strong research and development lab as for six months. Her current research interests include the analysis of Open Innovation processes, and internet-based collaboration platforms for project management.

Anna De Liddo is a Research Associate at the UK Open University's Knowledge Media Institute. Her research focuses on the socio-technical factors influencing the design and uptake of Online Deliberation and Collective Intelligence infrastructures. These are online systems which seek to improve collective awareness of the changing environment, and collective capacity to make sense of complex issues, particularly with regard to collaborative decision-making processes in the Environmental and Urban Planning domain. She has a particular interest in knowledge construction through discourse, and the role of technology in scaffolding dialogue and argumentation in contested domains.

Simon Buckingham Shum is Professor of Learning Informatics at the UK Open University's Knowledge Media Institute, and a Visiting Fellow at the University of Bristol Graduate School of Education. He researches, teaches and consults on learning analytics, collective intelligence and dialogue/argument visualization. He is a founding member of the Society for Learning Analytics Research, the Compendium Institute and LearningEmergence.net.