Supporting Group Decisions by Mediating Deliberation to Improve Information Pooling

Joshua E. Introne
Center for Collective Intelligence
MIT Sloan School of Management
Cambridge, MA USA
(617) 253-3566
jintrone@mit.edu

ABSTRACT

Group decision support systems (GDSS) hold significant potential for improving decision making, but they have not been broadly adopted. One reason for this is that these platforms introduce representational work for users that is distinct from a more familiar deliberative interaction but they offer uncertain payoff. This article presents a study with a platform that addresses this problem by leveraging the argumentative structure of deliberative conversation to drive a decision support algorithm. The platform uses argument visualization to mediate the collaborators' conversation. The study demonstrates that the platform addresses a known deficiency in human information pooling called the "common knowledge" phenomenon.

Categories and Subject Descriptors

H.5.3 [Information interfaces and presentation]: Group and Organization Interfaces---Computer-supported cooperative work, evaluation/methodology, synchronous interaction, web-based interaction; H.4.2 [Information systems applications]: Types of Systems---Decision Support

General Terms

Design, Experimentation, Human Factors

Keywords

Collaborative argument visualization, group decision support

1. INTRODUCTION

Important organizational decisions are often entrusted to groups instead of individuals because groups can access a larger and more diverse pool of information and expertise than individuals alone [17][23]. However, the deliberative exchange of information in groups is well known to be far from perfect, leading to documented problems such as groupthink [15], polarization [19] and biased information pooling [27].

Group decision support systems (GDSSs) may be able to address some of these problems, but they have not been broadly adopted in the workplace. One reason that has been offered for this is that the decision models used by these systems are too complex and make decision problems appear harder to stakeholders than they might otherwise [17] . Another possible explanation is that it is unclear exactly what GDSS offers to users[1]. This article seeks to address both of these problems.

To understand what makes GDSS "too complex" it is useful to consider those features such systems have in common. All GDSSs support parallel, asynchronous, persistent, possibly anonymized communication. These features can alleviate some of the process losses that plague collaborative discussions, for example by reducing social pressures that might otherwise deter people from voicing unpopular opinions, and relaxing the requirement that only one person speak at a time [20]. However, these features have not been found to eliminate the information processing problems found in face-to-face groups (e.g. [5]).

Thus, to improve group information processing, most GDSS systems also incorporate some sort of direct support for decision analysis. In general, these systems guide their users to construct a model of a decision problem that the system can evaluate algorithmically. For example, the Decision Lens™ platform (a commercial platform) requires users to decompose a decision problem into a set of options and set of criteria along which these options are to be compared. Participants then individually rank each of these criteria and options with respect to each criterion in pairwise fashion. The system uses an eigenvector analysis to measure the consistency of the rankings, and ultimately, to establish an overall ranking based on these pairwise comparisons. The system is then able to report back to the users which of the options best satisfies their collective priorities, as well as the degree of consistency (or consensus) among the users.

Most modern GDSS systems embody a decision process similar to that used by the Decision Lens™ platform. First, the decision problem is decomposed into a set of criteria that are important to stakeholders. These criteria are then ranked in some manner (e.g. by voting). Finally, alternatives are assessed according to these criteria, and results analyzed. However, the way this process functions with the collaborators' deliberative process varies between platforms. In Decision Lens™, this process is designed to replace a significant portion of the users' deliberative interaction. In other systems, such as GroupSystem's ThinkTank™, the model is more of a book-keeping mechanism that people use while they deliberate to keep track of things. In all cases, though, GDSS system introduce the decision model as an

GROUP'09, May 10–13, 2009, Sanibel Island, Florida, USA.
Copyright 2009 ACM 978-1-60558-500-0/09/05...$5.00.

artifact to be jointly created by the users so that this model may be analyzed mathematically.

GDSS systems thus fundamentally alter the nature of the work people do to make group decisions. If, as with the Decision Lens™, the model is designed to replace the deliberative process entirely, decision stakeholders must put their faith in an esoteric mathematical model. If the model is designed as an artifact to be used alongside deliberation, collaborators are required to manage both the task of coordinating their deliberative activity and constructing a model.

Asking any group of users to change a deeply engrained work practice (e.g. deliberative decision-making) is likely to be met with some resistance, but such changes can be worthwhile for organizations if there is a sufficient payoff. Unfortunately, there is little empirical data that demonstrates conclusively that GDSSs improve decisions in general. The empirical data that does exist can be hard to draw together to make any conclusive claims (compare [8] to [7]), and would-be end-users may be left wondering whether the costs of imposing a new decision process on an organization are worth it [1].

In this article, I seek to address the above concerns in the following manner. First, I present a system that sidesteps the problems that are associated with objectification of the decision model by leveraging structure present in the deliberative activity of decision makers to construct an underlying decision model, and uses this model to adapt the decision making process itself. I will then describe a study with the designed platform that demonstrates that the system is able to address a well-known problem with group decision-making called the "common knowledge" phenomenon.

This study is a preliminary investigation into a reconception of GDSS as mediated deliberation rather than collaborative interaction with a decision model. I am not concerned with the direct question of which approach is better. I merely seek to demonstrate first, that collaborators can use a deliberative formalism that can be leveraged to drive a decision model, and second, that such a model can be used to improve an aspect of decision making in a concrete manner.

The discussion proceeds as follows. First, I provide a very brief background on collaborative argument visualization, and describe how common approaches to argument visualization may be tied to a belief aggregation technique. I will introduce the GDSS platform that is the focus of discussion. Following a description of the system, I will describe the common knowledge phenomenon and present an emprical study that illustrates that the system addresses the problem. Finally, I will discuss my findings and offer some suggestions for future designs that build on these results.

2. ARGUMENT VISUALIZATION

Computer supported argument visualization (CSAV; see [2] for an excellent introduction) is a technique used to scaffold complex deliberations. These systems are usually based upon argument formalisms that are derived from either Toulmin's [31] theory of argumentation or Kunz and Rittel's [16] IBIS technique for structured dialog about "wicked" problems.

Toulmin-based argument visualization platforms such as Belvedere and Rationale have been thoroughly studied in learning environments. They have been shown to be highly usable and have been demonstrated to improve learning and collaborative knowledge creation along many dimensions (e.g. [10,30]). While they have not been broadly studied as decision-aiding platforms per se, they are beginning to be applied more frequently in such contexts (e.g. [10]).

In the best cases, argument visualization offers two features that are critical to designing a new kind of of GDSS platform. First, the structure it adds to deliberation is, by most accounts, work that users are comfortable with. Furthemore, it lends a degree of structure to deliberative conversation that might be leveraged by an algorithmic decision model. The potential for argument visualization to be used in this manner is discussed in [3], and Introne & Alterman [13] describe how this design approach can be used to introduce algorithmic support that improves collaborative activity in the general case.

Platforms like Rationale and Belvedere share a common formalism, shown in Figure 1. "Alternatives" represent options that are possible choices in a deliberation, and "arguments" support (pro) or refute (con) alternatives. Arguments may also support or refute one another. Das [4] describes how belief aggregation techniques like Dempster-Shafer theory [22] may be used in combination with a very similar formalism to support reasoning by autonomous decision agents. The formalism offered by Das does not explicitly discuss the chaining of arguments, but this may be accomplished if each argument is considered to be a choice between the argument itself and its antithesis.

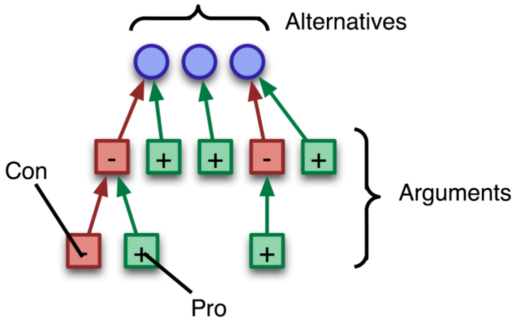
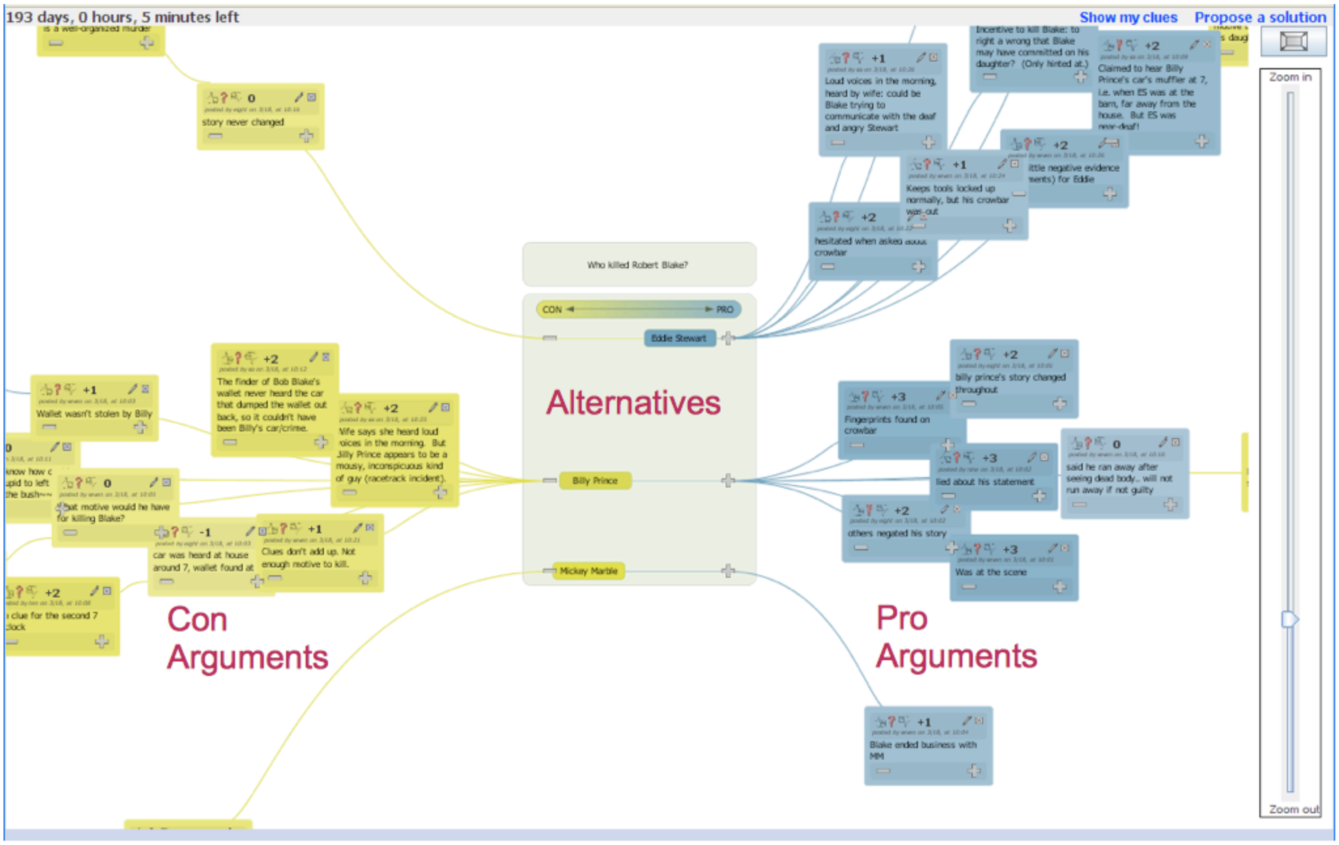


Figure 1: A common argument formalism

In this manner, argument visualization techniques that have previously been demonstrated to support the deliberative process may be used to gather information about a decision problem in a manner that can be used by an algorithmic belief aggregation technique. In the following section, I describe a platform that was implemented based on these ideas.

2.1 The Design of REASON

A platform called REASON (Rapid Evidence Aggregation Supporting Optimal Negotiation) was designed based on the insights in the previous section. The platform runs as a web application. The server handles all incoming requests, which read from or write to a domain model which embodies the argument formalism shown in Figure 1. At runtime, the domain model itself is a cached version of information that is maintained in the database. Any modifications to the data in the domain model cause the argumentation engine (which employs Dempster-Shafer theory over an argument network, as discussed above) to re-aggregate all information and the database to be updated. The logic that maps the domain model to the database is handled via



193 days, 0 hours, 5 minutes left

Show my clues Propose a solution

Who killed Robert Blake?

CON PRO

Eddie Stewart

Billy Prince

Mickey Marble

Alternatives

Con Arguments

Pro Arguments

is a well-organized murder

0 story never changed

+1 Wallet wasn't stolen by Billy

+2 The finder of Bob Blake's wallet never heard the car that dumped the wallet out back, so it couldn't have been Billy's car/crime.

+2 Wife says she heard loud voices in the morning. But Billy Prince appears to be a mousy, inconspicuous kind of guy (racetrack incident).

0 What motive would he have for killing Blake?

-1 car was heard at house around 7, wallet found at

+1 Clues don't add up. Not enough motive to kill.

+2 clue for the second 7 o'clock

+1 Loud voices in the morning, heard by wife: could be Blake trying to communicate with the deaf and angry Stewart

Incentive to kill Blake: to right a wrong that Blake may have committed on his daughter? (Only hinted at.)

+2 Claimed to hear Billy Prince's car's muffler at 7, i.e. when ES was at the barn, far away from the house. But ES was near-deaf!

+2 little negative evidence (arguments) for Eddie

+1 Keeps tools locked up normally, but his crowbar was out

+2 hesitated when asked about crowbar

+2 billy prince's story changed throughout

+3 Fingerprints found on crowbar

+3 lied about his statement

0 said he ran away after seeing dead body.. will not run away if not guilty

+2 others negated his story

+3 Was at the scene

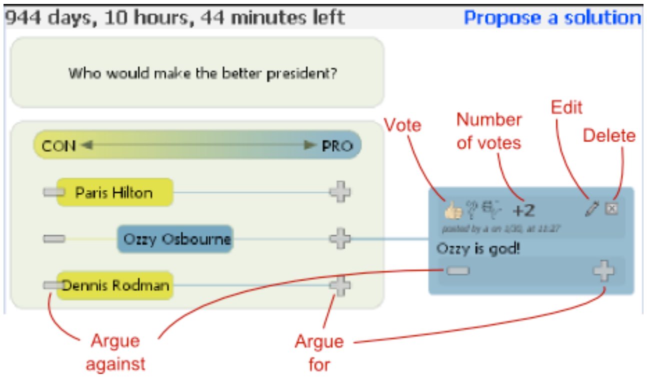
+1 Blake ended business with MM

Figure 2: The REASON interface

an object-relational mapping (ORM) layer, implemented using an open source product called Cayenne.

The interface is implemented as a Java™ applet that runs in most modern browsers. The applet polls the server continuously, so any updates propagate to all connected users within the update interval. The interface presents users with a graph-based visualization of the domain model. The interface is built using the Prefuse (http://prefuse.org) library, an open-source visualization toolkit. A screenshot of the system is shown in Figure 2. Alternatives are represented by "bubbles" in the box in the center of the display. These central bubbles drift to the right (pro) or left (con) depending upon how much relative weight is assigned by the aggregation engine. Arguments are represented by nodes in subtrees that are attached to the alternative. Arguments are colored yellow if they disagree with their parent, and blue if they agree. The initial argument for or against an alternative determines whether a subtree extends to the right (pro) or left (con) of the alternative. The user can zoom in and out, pan the display, and automatically re-center and fit the graph to the window.

The nodes of the graph in the visualization "float," and are automatically laid out via an animated force-directed layout algorithm, such that nodes exert a repulsive force, and links exert spring force. Force planes are employed to keep nodes in different subtrees separate. Because the graph is continuously animated, users are able to drag individual nodes, and this will pull attached nodes along with the dragged node – upon releasing the node, all nodes will drift back to a position determined by the layout algorithm. Clicking on any given node will select that node, its ancestors (up to an alternative), and its descendants. Selected nodes are zoomed in, and the animation for those nodes is paused, so that the user can control their placement. Other nodes (which are not part of the selected set) are de-emphasized and remain animated.



944 days, 10 hours, 44 minutes left

Propose a solution

Who would make the better president?

CON PRO

Paris Hilton

Ozzy Osbourne

Dennis Rodman

+2

posted by a on 1/30, at 11:27

Ozzy is god!

Figure 3: Reason Node Detail

All message posting and voting occurs via the graph. Controls are included in each existing argument and for each alternative to allow the user to post a new argument for or against that item (see Figure 3). Once an argument has been posted, the user can change their vote (in favor, against, or neutral). The username of the posting user, the time when the post was made, the user's current vote, and the sum of all votes assigned to that argument are all displayed (note that because votes may be either "up" or "down" this number can be negative). The user may also edit the text or valence of the post, or delete it. Deletion will cascade to all children.

Belief aggregation might potentially be used in many ways, but for the sake of the current study, it is used as follows. First, users are provided with a continuous display of the "winning" alternative (reflected by the relative horizontal positions of the alternatives) according to the belief aggregation algorithm. Users might use this feedback to either argue a point more strenuously by posting new arguments, or perhaps to re-evaluate their own opinions. However, it is possible for users to ignore this feedback. To guarantee that the system actually mediates the decision process, any final consensus must match the winning alternative as determined by the system. Thus, users may click "propose a solution" at any point, and this will initiate a vote on the system's currently assessed solution. If the outcome of a vote is not unanimous, collaborators continue to deliberate.

As discussed in the introduction, REASON does not present the model to its users as an objectified artifact to be constructed in support of decision making. Rather, collaborators deliberate *though* the system, and a byproduct of this deliberation is the construction of the decision model that can be evaluated to adapt the group decision-making process in a variety of ways.

To demonstrate that the system is capable of improving decision-making in a concrete manner, groups of decision makers were examined under conditions that have been shown to lead to a specific group information processing problem referred to as the common knowledge phenomenon. The following section introduces the common knowledge phenomenon, and then an empirical study is described.

3. THE COMMON KNOWLEDGE PHENOMENON

A persistent problem with small group decision-making is the inability of a group to pool and process all the information available to its constituent members. Groups tend to focus their discussions preferentially upon information that members hold in common prior to group interaction. In addition, pre-discussion biases (biases that individuals have prior to group discussion) have substantial influence on decision outcomes, "as if group members exchanged and combined their opinions but paid little attention to anything else" ([12]; pg.132). This problem has been referred to as the "common knowledge" phenomenon [11]. The primary vehicle for examining the common knowledge phenomenon is the *hidden profile* experiment, first introduced by Stasser and Titus [27].

In a hidden profile experiment, members of a group are asked to make a choice between several alternatives after pooling their information. Information is distributed among participants so that some individuals have some information that others don't, and the correct choice can only be determined by considering all of the information. For instance, consider a decision task with two options (A and B) and three collaborators (P1, P2, P3). Information shared by the participants is referred to as "shared" information, and information held by only one of the participants is referred to as "unshared" information.

As shown in Table 1 shared information contains two distinct pieces of information supporting option A, and one piece of information against B. Thus, shared information recommends option A. Furthermore, combining each individual's unshared information with the shared information leads each individual to favor A prior to discussion. However, if all shared and unshared information is considered together, decision option B has the most support. This is the typical structure of hidden profile studies, and it allows psychologists to study the effectiveness of information pooling in group decision-making.

Table 1: A typical distribution of information in a hidden profile task

	P1	P2	P3
Shared	2A+,1B-		
Unshared	1A-,1B+	1A-,1B+	1A-,1B+

In their seminal paper, Stasser & Titus [29] found that groups engaged in a hidden profile task were less likely to retrieve and discuss unshared information and hence less likely to identify the hidden profile (the correct solution). Instead, a group's decision was best predicted by the distribution of pre-discussion preferences, and was typically consistent with shared information. Numerous manipulations of experimental variables and more precise analyses have elaborated upon this finding, but the basic phenomenon remains fairly consistent across these studies [17,30,34].

Several theories have been offered as explanations for the common knowledge phenomenon. Among these theories there is disagreement about whether the cause lies with a group's retrieval and exchange of information or with the way a group combines that information to arrive at a decision. Stasser & Titus [27] suggested that because more people have been exposed to the information that is shared, it is more likely to be recalled and discussed in conversation.

However, some studies with simple GDSS platforms (i.e. offering only basic messaging capabilities) have shown that decision-making is still biased towards common information even when when problems related to biased information retrieval and exchange are addressed [5,6,18,24]**.** This suggests that even though biased information retrieval and exchange in groups might offer a partial explanation for the common knowledge phenomenon, it is not the whole story. How people combine their information to arrive at a final decision must also be a component of the common knowledge phenomenon.

The hypothesis underlying the user study described in the following section is that REASON can fix some of the problems with the way people combine their information by countering the excessive weight placed on common information.

4. USER STUDY

The primary objective of the user study was to determine if REASON could address the information pooling problem that is implicated in the common knowledge phenomenon. In order to do this, a second system was generated to serve as a control condition. This system did not employ belief aggregation, but was otherwise identical to the system described in the previous section. The same argument formalism was used, but the system did not display the weight for the alternatives to the users, and final decision proposals were left up to the users. For ease of reference, this platform will be referred to as the non-aggregating, or "NA" platform. Accordingly, the platform equipped with belief aggregation will be referred to as the "A" platform.

Because the goal of the study was to examine the platform as a means for addressing problems observed under hidden profile conditions, the study design closely follows previous work with hidden profiles. The experiment materials themselves are a slightly modified version of those used in several previous hidden profile investigations (e.g. [9,25,26]). The details of the study design are described in the following sections.

4.1 Design

The study was run as a single factor experiment, where groups used a version of REASON that either had aggregation (A) or did not have aggregation (NA).

115 university students were recruited for the study. Participants were randomly assigned to each condition, based upon when they were available. Each was paid twenty dollars for their time, and promised a free movie ticket if they chose the correct suspect. Subjects were not told if they had "won" until after all 115 participants had been run through the experiment.

Table 2: Distribution of clues in the study; + indicates an implicating clue, - indicates an exonerating clue

	E	B	M
Shared	+3	+6	+6
Unshared	+3	-3	-3
Total Weight	+6	+3	+3

Groups of five were used in the experiment. This number was chosen as representative of small group decision-making in "real-world" scenarios, and as the smallest size group that might begin to see significant process gains due to GDSS support [8]. Of the twenty-three groups that used the system, three were omitted from analysis, due to data collection errors (two groups) and technical problems with the application (one group). The remaining twenty groups were split evenly between the two conditions.

4.2 Decision Task

The task is a fictional homicide investigation. The mystery involves three suspects, E, B, and M. It was presented to the subjects online as a set of affidavits and supporting information. The mystery is identical to the one used in [25] with the following modifications. In a pilot study, it was discovered that two of the clues offered one of the suspects (M) an air-tight alibi. These clues were modified in an attempt to balance the fact pattern. Additionally, names of characters in the mystery were modified when it was found that the mystery could be easily found online.

The murder mystery contains twenty-four clues that either implicate or exonerate each of three suspects. These clues are organized as shown in Table 2. Information was distributed among the participants so that three of the members had some unshared information, and the other two had only shared information. Participants with unshared information were given unshared clues about only one of the suspects. Thus, the "E" expert was given the three unshared clues implicating E in addition to the shared information, and so forth. In the following, the "E" expert will be referred to as EE, the "B" expert BE, and the "M" expert ME. The participants with only shared information will be referred to as S1 and S2. This distribution implies the following pre-discussion preferences: EE should be equally likely to choose any of the three options; BE should choose option M; ME should choose option B; and S1 and S2 should both be equally likely to choose either B or M. There should be no statistical pre-discussion bias supporting any single option, but there is a bias away from the correct option (E).

4.3 Procedure

A pilot study was run in which users were allowed to work asynchronously over the course of a week, but participation was too uneven for this to be a viable strategy; only two of eight groups studied in the pilot were able to achieve consensus by the end of the allotted time. Thus, the full experiment was run synchronously, in a lab. In the laboratory setting, participants were not allowed to talk to one another during the collaborative portion of the experiment, could not easily see each other's screens, and were assigned anonymous names.

The entire experiment was implemented as a timed web-application. The phases of the experiment included an introduction, a consent form, a timed training application with instructions, a timed pre-study period during which participants read the mystery materials, a timed collaboration period, and an exit interview. The training application was identical to the application that would be used during the collaborative period, but an unrelated decision problem was used ("Who would make the best president?"). Each timed period was assigned a deadline, and deadlines were 30 minutes apart. Thus, if all participants finished the pre-study early, additional time was added to the collaborative portion of the task. All participants were required to finish with the pre-study before the collaborative task could begin. In practice, most groups took the entire time allotted to each of the timed periods.

Participants were told that information was distributed unevenly and that they would need to pool all of their information to identify the correct option. The mystery was presented as a set of affidavits (and a few other pieces of evidence), and the clues were not highlighted or identified for the participants in any way. The mystery materials were available both during the pre-study, and during the collaborative problem period.

Participants were given some very simple guidelines for using the argumentation software. In particular, they were told to post short single-clause statements that did not include conjunctions like "and," "or," or "if." They were also told that new clues should be posted at the "top" level, next to the alternative they addressed, and that discussion of these clues should occur in the threads attached to these clues. Finally, they were told that clues about a given alternative should only be posted in the thread attached to that alternative.

A variety of information was collected throughout the study. Each deliberation was stored in a database for subsequent analysis. All data from submitted forms (consent, pre-discussion opinions, and exit survey) was also collected in a database. A substantial amount of information was also collected in the web log. The logged data included selection activities in the interface (which were explicitly logged) and all timing information.

5. RESULTS

The twenty groups generated a total of 1146 posts (arguments), split almost evenly between the two conditions (604 for the NA groups, and 542 for the A groups). On average, the A groups authored 60.2 posts (*SD=12.8*) and NA groups authored 67.1 posts (*SD=18.9*). This difference was not significant. The entire corpus of collected data is roughly 14k words.

5.1 Pre-discussion Choices

Despite efforts to balance the fact pattern (via the replacement of one clue, as discussed above), an examination of pre-discussion choices revealed that the information contained in the mystery was uniformly biased away from suspect M (see Table 3). Similar results are not reported in the prior work using this mystery [25,26,28], so there is no basis for comparison.

Table 3: Pre-discussion opinions for each participant.

Participant	B	E	M
EE	.4	.45	.15
BE	.35	.4	.25
ME	.7	.2	.1
S1+S2	.675	.175	.15

Nonetheless, the evidence manipulation did have an impact on individual pre-discussion choices consistent with the intended impact, although there was a weaker preference for M than expected for all evidence sets. A two-way ANOVA over experimental condition and evidence set demonstrated significant differences between the evidence sets (*F(2,24)=5.23, p<.02*), but no differences between the experimental conditions, and no interactions between experimental condition and evidence set.

5.2 Discussion Content

To determine which outcome was supported by exchanged information, it was necessary to examine this information to determine which clues were discussed. The entire corpus was tagged to identify where clues appeared. A clue was considered to be in a post if the specific piece of information is mentioned, and every instance of a clue was documented for each post. The coding of clues has not yet been validated with independent coders, but was unambiguous and it is not anticipated that validation studies will substantially alter the described results. In the following, I will refer to the property of a clue being shared or not as *the clue-class*, and the suspect the clue is about as the *clue-suspect*.

Table 4: An example of a valence error

Line	Valence	Post
1	[Top level exonerates M]	According to the detective's timeline, Mickey didn't have time to kill Blake and get to the golf course when he did.
2	CON	Do we have proof of when he got to the golf course, or is it hearsay?
3	CON (wrong)	I think we only have him saying it.

In analyzing the data, it was observed that several posts were of the wrong argument type – e.g. a "con" was used when in fact the content of the argument reflected a "pro." These errors will be referred to as valence errors, after the psychological definition of valence meaning the attractiveness or aversiveness of a situation.

Valence errors occurred frequently after questions, which are not part of the formalism. For example, in Table 4, the individual in line (2) asks whether or not there is any proof of M's alibi. This is a negative leaning question, and has been correctly posted as a rebuttal to the statement in line (1). However, while the post in line (3) would seem to agree with the argumentative force of line (2), it is posted as a rebuttal to line (2). This is apparently a response to the first clause of the question in line (2), and so makes some sense, but is technically incorrect with respect to the argument. The above example illustrates one of the difficulties users experienced with the formalism.

Table 5: Summary statistics for the experiment

	A	NA	p-value
# Threads	26.4 (*sd. 8.69*)	19.7 (*sd 4.76*)	*P<.05*
Thread Size	2.34 (*sd 2.01*)	3.82 (*sd. 4.1*)	*P<.001*
Valence Error Rate	.04 (*sd .03*)	.10 (*sd .06*)	*P<.01*
# Clue Threads	.48 (*sd.5*)	.51 (*sd .5*)	--
# Clue Mentions / Thread	.81 (*sd .35*)	.83 (*sd .16*)	--
Thread Clue Density	.37 (*sd .13*)	.26 (*sd .1*)	*P<.05*

Despite these difficulties, users performed quite well considering the limitations of the formalism. Table 5 contains statistics about the frequency of valence errors. Note that participants in the NA condition (without belief aggregation) were significantly more likely to make valence errors than those in the A condition. The rate is less than one in ten across all arguments, though, and this provides a rough indicator that collaborators can indeed use the formalism with minimal training.

Table 5 also provides several additional summary statistics for the collected data. These statistics indicate some general differences in the overall behavior of interlocutors across the two conditions. In this domain, a thread is considered to contain all arguments beneath and including an argument that is connected to an alternative. Loosely speaking, a thread may be thought of as discussion about a particular piece of evidence in relation to a suspect, although threads often covered multiple pieces of evidence.

Participants in either condition had roughly the same rate of discussion for each type of clue. Figure 4 shows the relative proportions of clues mentioned by participants for each of the six categories (shared or unshared for each of the three suspects) of clues. The graph reflects the average proportion of total possible clues in each category that were mentioned at least once during a discussion. In general, unshared information is mentioned less in conversations than shared information. A three-way ANOVA (clue-class × clue-suspect × experimental condition) revealed a significant main effect for clue-class (whether a clue was shared or unshared) (*F(1,36)=6.63, p<.02*), but no difference across experimental conditions, and no interactions.

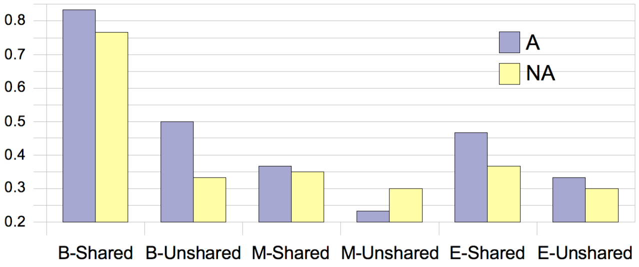


Figure 4: Relative proportions of clues mentioned by each of the participants.

As is apparent in Figure 4 there was also a clear difference in the proportion of clues mentioned for each of the three suspects.

The above analysis confirmed this, revealing a highly significant main effect for clue-suspect (*F(2,36)=8.71, p< .005*). This can be explained by the distribution of pre-discussion preferences. If shared and unshared clues are combined for each suspect, there is perfect correlation in each condition between the average number of clues mentioned for each suspect and overall user pre-discussion preferences *(R=1.0* for both conditions).

A similar analysis was performed on clue-repeat rate (the average number of times a clue was repeated). Once again, a three-way ANOVA (clue-class × clue-suspect × experimental condition) revealed a highly significant main effect for clue-suspect *(F(2,36)=9.45, p<.001*), and once again the overall rate for each suspect correlates very highly with the pre-discussion opinions of the users (*R=.98* for the adaptive case, and *R=1.0* for the non-adaptive case). There was not, however, a main effect for experimental condition.

In summary, the above analyses confirm prior findings with respect to the common knowledge problem. Participants were significantly more likely to mention shared clues than unshared clues across the two conditions. Furthermore, this effect is heavily mediated by pre-discussion opinion, which appears to dictate not only the likelihood a given clue will be mentioned, but also how frequently it is mentioned.

5.3 Information Pooling

The primary hypothesis established at the outset of the case study was that the system would improve participants' ability to pool the information they exchanged and make decisions that are consistent with that information. Demonstration of this hypothesis indicates that the platform helps people avoid over-weighting information that is shared. The results described below confirm this hypothesis.

Decision quality was assessed by determining whether or not the clues that were exchanged support the decision that is ultimately made by the group. Following the design employed in [26], each clue is given a unit of weight, and its valence (implicating or exonerating) determines its sign (+ or -, respectively). Weights for each clue that is mentioned in the discussion are then added together, and the suspect with the most weight is determined to be the best supported suspect. In case of ties, the group decision is considered to be "consistent" if it is among the best supported options.

Table 6: Information pooling results for the two conditions

	A	NA
Consistent (Unique)	4	1
Consistent (Two options)	2	0
No Decision	1	2
Inconsistent	3	7

Results of this analysis are shown in Table 6. A two-tailed, unpaired T-test reveals that the A group's decisions were significantly more consistent with mentioned clues than the NA groups (*p<.02*). Thus the A groups were likely to make joint decisions based on the clues they exchanged in conversation, and the NA groups were not. This result establishes a strong correlation between information pooling behavior and the type of platform used. Thus, at this level of analysis, the features supported by the decision model in the A platform appear to transform the collaborative process so that it is closer to the "rational" ideal embodied in the system. The next section establishes a stronger causal relationship between platform use and this outcome.

5.4 System Use

To make a stronger case for a causative relationship between the A platform and group information pooling performance, we would like to be able to say that people used their available information correctly, represented it with the provided formalism as intended, and that the information represented in this manner was responsible for the final outcome. Several questions should be asked. First, was conversation about clues represented within the provided formalism correctly? If so, this indicates that participants assessed the information content of each clue as intended (validating the experimental setup), and that they were also able to use the formalism to encode the value of these clues (validating the platform). If clue weights assessed by the system do not correlate well with the way the clues are used by the interlocutors, it is important to understand whether the problem was with the interpretation of the clues or with use of the formalism.

Additionally, we need to verify that it is the information about clues that is responsible for the outcome. Does the conversation focus on clue information, or other information that is not considered to be a clue in the experiment materials? How are the two types of conversation related? If non-clue discussion explains the outcome better than clue discussion, it may indicate that the experimental design is flawed because information that was not considered to be relevant in the experiment design was in fact relevant to participants. However it might also indicate the need for extensions to the platform to control for such "non-informational" activity. The platform as designed assumes that people are able to focus primarily upon what is important in a decision, or at least that the weight of what is important correlates with what is not in conversation.

To answer the above questions, conversation "about" clues (henceforth "clue discussion") was segregated from other conversation ("non-clue discussion"). The weights contributed by these portions of the conversation could then be compared to the overall weight of the conversation and the weights calculated just according to the exchanged clues (using Pearson's R).

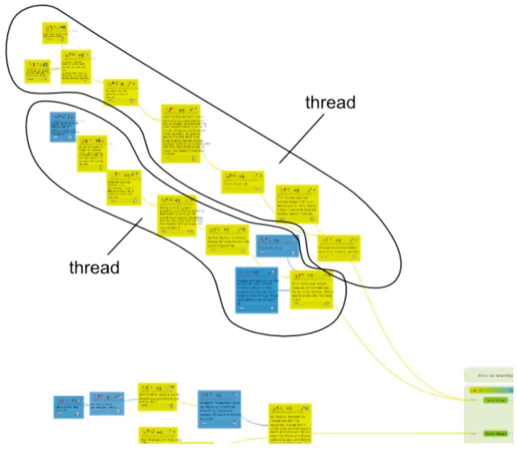


Figure 5: Threads in an argument

To segregate the conversation in this manner, discussion following the mention of a clue in a thread was considered to be clue discussion about that clue. For the sake of this analysis, a "thread" is considered to be all discussion connected to a "top-level" argument (see Figure 5).

An example of "clue" discussion that follows "non-clue" discussion is shown in Table 7. Although there are examples where it seemed that non-clue discussion follows clue discussion, the approach worked well in the "A" case upon inspection. Furthermore, segregating conversation in this manner sidesteps the difficulty of defining a repeatable subjective criteria for determining whether or not a section of dialog is "about" a clue. Generally, the consistency of results obtained suggested that this was a reasonable approach in the "A" case. The "NA" case was a somewhat different story, and is discussed below.

Table 7: Clue discussion following non-clue discussion

Type	Post
Non-clue	1. (ME) I think it was either Mr. Marble or Mr. Stewart.
Non-clue	2. (BE) so how do people stand on this? Aka which one do you think is guilty?
Non-clue	3. (ME) I think Eddie is guilty but I'm still looking for a sufficient motive, probably with his daughter.
Clue (Billy's fingerprints on crowbar)	4. (S2) I agree with this, but I want some theory on how he got billy's prints on the crowbar.
Clue (Discussion about Billy's fingerprints).	5. (ME) Good point, I think it's possible that he removed it from the scene of the crime without thinking for some reason.

5.4.1 The "A" Case

Three of the six groups with consistent outcomes in the "A" case performed as expected. In each of these groups, the weight contributed by "clue-discussion" was positively correlated with both the clues ($R>.75$) themselves, and the total overall weight as assessed by the system ($R>.75$). The other three consistent groups had somewhat different usage patterns. In each of these cases it was possible to determine why this deviation occurred. One group made a single valence error that dramatically reduced the correlation between clue-discussion and the clues themselves (from $R=.72$ to $R=.04$). Another group consistently evaluated implicating clues for B incorrectly, in line with that group's strong pre-discussion bias for E, also leading to a low correlation between clue-discussion and the clues themselves. The remaining group had one member (the EE member) who did not bring any clues to the discussion. As a result, the conversation was dominated by non-clue discussion, which was more highly correlated with total weight ($R=.9$) than was clue discussion ($R=.2$). However, the non-clue discussion embodied a correct reasoning process even though it did not contain actual clues.

Each of the four remaining groups that did not make a consistent decision failed to do so for distinct reasons, as follows:

1. ***Gaming the system*** – One group explicitly overturned the system's evaluation at the end of the conversation in order make the system agree with the group's consensus.

2. ***Poor information retrieval*** – One group failed because it in general did a very poor job retrieving any information relevant to the mystery and relied instead on character assessments.

3. ***Contentious disagreement*** – In one group, two individuals who strongly disagreed engaged in a voting "war." Together, these two individuals were responsible for 81% of all votes contributed by the group, and this skewed the system's assessment.

4. ***Incomplete discussion*** – The one group that did not achieve consensus only retrieved one unshared clue, and non-clue discussion dominated the conversation. Informally, the group seemed never to come to consensus around a story that might explain events.

In summary, groups in the A case were generally able to use the system as intended. However, it is clear that there are several failure modes that can limit the effectiveness of the platform. Some of these problems, like poor information retrieval or incorrect interpretation of clues, are not so much failures of the system or approach as they are limitations of human information processing. Other problems, however, like gaming the system and incorrect valence posting, offer clear avenues for future extensions. Several design suggestions based on these problems are mentioned below.

5.4.2 The "NA" case

A similar analysis was performed for groups in the NA case. However, it turned out that there was very little to be gained from such an analysis. The seven cases where groups made inconsistent decisions may be initially divided into two classes. In three groups, the aggregate assessment that would have been made by the system matched the distribution of clues discussed, but the group made a different decision. In the other four groups, the system's assessment matched the group's choice, but this did not match the distribution of clues. Within these two classes there was very little regularity in how the different kinds of discussion (clue vs. non-clue) correlated with either the clues themselves or the total weight assessed by the system.

One explanation for the apparent lack of pattern across the distribution of weights from different portions of the conversation is that the approach used to segregate clue-chat from non-clue chat was not valid in the NA case. Several indicators support this explanation.

As illustrated above (Table 5), participants in the non-adaptive case make more valence mistakes and create longer threads that are not as densely populated with clues. There are also significantly fewer threads per conversation. Further investigation also revealed that participants in the NA condition were significantly more likely to argue both sides of an issue in any given thread. 35% of threads in the non-adaptive case contain at least one instance of a participant posting an argument counter to their initial position. In the adaptive case, this behavior occurs in 20% of all threads. This difference is highly significant ($p<.001$).

Thus, people in the A condition used threads differently than did people in the NA condition. In the A case, collaborators create short, clue-dense threads. In the NA case, collaborators have longer, more diplomatic conversations, perhaps covering more topics. Because of this difference, the assumption that a thread is "about" a clue for its remainder once a clue has appeared may not be valid.

The preceding data and analysis establish that the clues exchanged by participants in the NA groups in conversation were not as predictive of the groups' final answers as the clues exchanged by the A groups, and that the NA groups had qualitatively different conversations. However, despite these differences, the NA groups got the "correct" answer nearly as often (50%) as groups in the A condition (60%). This finding led to a deeper analysis.

Although full presentation of this deeper analysis is outside of the scope of this article, it has led to the following findings (discussed in [14], pp. 204-262). First, the dynamics of the conversation differed markedly between the two groups. Seven of the ten NA groups had conversations that exhibited one or several "peaks" of activity. These peaks generally appeared during the last half of the conversation. During these peaks, most of the participants posted to a single thread, posts appeared quickly, and collaborators agreed with one another. These peaks of shared, focused agreement were completely absent in eight of the ten A groups, and less pronounced in the two A groups where they appeared.

A further preliminary finding is that the conversation during these peaks of activity focused upon constructing a narrative that could explain a handful of the clues that had been discussed. The term "narrative" is used here in the same sense as described in [21]. The narratives discussed during peaks of shared, focused agreement became the basis for these groups' final decisions. Again, these periods of shared story creation were almost non-existent in the A groups.

In summary, the inclusion of belief aggregation in REASON and the mediated decision process based upon it substantially altered the decision making strategies used by users. On the one hand, the platform appeared to encourage a more "rational" information pooling process, helping collaborators to evaluate shared and unshared information more equitably and consequently addressing a piece of the common knowledge phenomenon. On the other hand, it seemed to disrupt a collaborative story formation process. These results have many implications, some of which are discussed below.

6. DISCUSSION & FUTURE WORK

Most modern GDSS systems attempt to scaffold group decision making in order to guide users in the construction and analysis of a decision model. REASON offers a somewhat different approach. In REASON, the collaborators' deliberative engagement becomes the driver for the creation of a decision model that can in turn be used to adapt the decision making process. The decision model is not a focus for the users.

This study presented in this article demonstrates that people can use the argument formalism embodied by REASON to deliberate effectively, and in doing so, create a decision model that may be evaluated algorithmically. As demonstrated, the decision model can in turn be used to encourage less biased information-pooling behavior.

The analysis of the system's use suggests a number of avenues for improvement. One difficulty with the system as implemented was that people occasionally made mistakes in selecting the valence of a post, which can have a very large effect on the system's assessment. Valence errors were observed to occur frequently in response to questions. Extending the formalism to include questions may help to eliminate these problems.

Another interesting observation is that two of the groups skewed the system's results by misusing the representation. One way to address this problem would be to require users to "ground" their arguments in known, trusted sources. However, the fact that people misused the system indicates that the representation that was ultimately constructed by the users did not reflect what they interpreted the overall meaning of their conversation to be. This may indicate that the system needs to be redesigned so that it is easier for collaborators to see why the system makes the assessment it does. It may also indicate that the argument formalism employed was not sufficient to represent domain specific information in a manner compelling to collaborators.

This latter observation is bolstered by the findings that collaborators in the NA groups did not use the system formalism as "well" as the A groups (more valence errors, multiple topics per thread), and also approached the mystery-solving task as a collaborative story creation task. This story creation process has been documented in jury-decision making [21], and may have been instrumental in the NA groups' ability to identify the correct answer despite less than accurate information processing. This suggests that the system might benefit from an explicit means for representing and reasoning about causal models.

At a deeper level, the data presented here also suggest a more nuanced interpretation of the common knowledge problem. The majority of hidden profile studies have chosen tasks in which individual pieces of information can be evaluated in isolation with respect to the decision. A common task is the candidate selection task, in which positive and negative attributes about each candidate are distributed among collaborators.

Among hidden profile studies, the murder mystery task is an anomaly in that the connections between the clues play a significant role. However, in "real world" decision problems, connections between attributes in decision problems may be more commonplace. With the exception of [9], few have explored the impact of such connections upon information pooling and decision-making in hidden profile tasks.

The results presented here suggest that when such connections exist, they play a significant role for decision makers. The NA groups did not simply make decisions based on the proportion of clues retrieved. Rather, they attempted to construct stories and fit the clues recalled to these stories. Shared clues were indeed recalled and repeated more frequently, but a story creation process heavily mediated the role these clues played. Hence, care must be taken before extending laboratory observations of the common knowledge problem to real world decision-making (as in Sunstein's *Infotopia* [29]).

This leads to a final observation regarding normative approaches to decision making. Traditionally, normative approaches to decision support employ a subjective expected utility criterion as an ideal decision model. At their core, REASON and other GDSS systems embody such a model. The results here demonstrate that such a model may have unintended consequences. Indeed, it seems that such a decision model may help people to decompose a problem and evaluate information more equitably. However, such an approach may disrupt a shared story formation process that can help people operate under conditions of incomplete information. Future approaches to group decision support may benefit from considering these issues.

7. ACKNOWLEDGMENTS

This work was supported by DARPA contract numbers W31P4Q-05-C-R038 and W31P4Q-06-C-0366.

8. REFERENCES

[1] Briggs, R.O. 2006. On theory-driven design and deployment of collaboration systems. *International Journal of Human-Computer Studies 64*, 7, 573-582.

[2] Buckingham Shum, S.J. 2003. The Roots of Computer Supported Argument Visualization. In P.A. Kirschner, S.J. Buckingham Shum and C.S. Carr, eds., *Visualizing argumentation: software tools for collaborative and educational sense-making*. Springer-Verlag, 3-24.

[3] Chklovski, T., Ratnakar, V., and Gil, Y. 2005. User interfaces with semi-formal representations: a study of designing argumentation structures. *Proceedings of the 10th international conference on Intelligent user interfaces*, ACM, 130-136.

[4] Das, S. 2005. Symbolic Argumentation for Decision Making under Uncertainty. *7th International Conference on Information Fusion, 2005.*

[5] Dennis, A.R. 1996. Information Exchange and Use in Group Decision Making: You Can Lead a Group to Information, but You Can't Make It Think. *MIS Quarterly 20*, 4, 433-457.

[6] Dennis, A.R., Hilmer, K.M., and Taylor, N.J. 1997 Information exchange and use in GSS and verbal group decision making: effects of minority influence. *Journal of Management Information Systems 14*, 3, 61-88.

[7] Fjermestad, J. 2004. An analysis of communication mode in group support systems research. *Decision Support Systems 37*, 2, 239-263.

[8] Fjermestad, J. and Hiltz, S.R., 1998. An assessment of group support systems experimental research: methodology and results. *Journal of Management Information Systems 15*, 3, 7-149.

[9] Fraidin, S.N. 2004. When is one head better than two? Interdependent information in group decision making*1. *Organizational Behavior and Human Decision Processes 93*, 2, 102-113.

[10] van Gelder, T. 2003. Enhancing Deliberation Through Computer Supported Visualization. In P.A. Kirschner, S.J. Buckingham Shum and C.S. Carr, eds., *Visualizing argumentation: software tools for collaborative and educational sense-making*. Springer-Verlag, 97-115.

[11] Gigone, D. and Hastie, R. 1993. The common knowledge effect: Information sharing and group judgment. *Journal of Personality and Social Psychology 65*, 5, 959-974.

[12] Gigone, D. and Hastie, R. 1997. The impact of information on small group choice. *Journal of personality and social psychology 72*, 1, 132-140.

[13] Introne, J. and Alterman, R. 2006. Using shared representations to improve coordination and intent inference. *User Modeling and User-Adapted Interaction 16*, 3, 249-280.

[14] Introne, J.E. 2008. Adaptive mediation in groupware. Doctoral Thesis. ProQuest/UMI Publication No: AAT 3319826. Brandeis University.

[15] Janis, I.L. 1982. *Groupthink*. Houghton Mifflin Boston.

[16] Kunz, W. and Rittel, H. 1970. *Issues as elements of information systems*. Center for Planning and Development Research, University of California at Berkeley.

[17] Limayem, M. and DeSanctis, G. 2000. Providing Decisional Guidance for Multicriteria Decision Making in Groups. *Info. Sys. Research 11*, 4, 386-401.

[18] Mennecke, B.E. 1997. Using group support systems to discover hidden profiles: an examination of the influence of group size and meeting structures on information sharing and decision quality. *International Journal of Human-Computer Studies 47*, 3, 387-405.

[19] Moscovici, S. and Zavalloni, M. 1969. The Group as a Polarizer of Attitudes. *Journal of Personality and Social Psychology 12*, 2, 125-135.

[20] Nunamaker, J.F., Dennis, A.R., Valacich, J.S., Vogel, D., and George, J.F. 1991. Electronic meeting systems. *Commun. ACM 34*, 7, 40-61.

[21] Pennington, N. 1981. Causal reasoning and decision making: the case of juror decisions.

[22] Shafer, G. and Logan, R. 1987. Implementing Dempster's rule for hierarchial evidence. *Artificial Intelligence 33*, 3, 271-298.

[23] Shaw, M.E. 1981. *Group Dynamics: The Psychology of Small Group Behavior*. McGraw-Hill Companies.

[24] Shirani, A.I. 2006. Sampling and pooling of decision-relevant information: Comparing the efficiency of face-to-face and GSS supported groups. *Information & Management 43*, 4, 521-529.

[25] Stasser, G. and Stewart, D. 1992. Discovery of hidden profiles by decision-making groups: Solving a problem versus making a judgment. *Journal of Personality and Social Psychology 63*, 3, 426-434.

[26] Stasser, G., Stewart, D.D., and Wittenbaum, G.M. 1995 Expert Roles and Information Exchange during Discussion: The Importance of Knowing Who Knows What. *Journal of Experimental Social Psychology 31*, 3, 244-265.

[27] Stasser, G. and Titus, W. 1985. Pooling of unshared information in group decision making: Biased information sampling during discussion. *Journal of Personality and Social Psychology 48*, 6, 1467-1478.

[28] Stewart, D.D. and Stasser, G. 1998. The sampling of critical, unshared informationin decision-making groups: the role of an informed minority. *European Journal of Social Psychology 28*, 95-113.

[29] Sunstein, C.R. 2006. *Infotopia: How Many Minds Produce Knowledge*. Oxford University Press, USA.

[30] Suthers, D.D., Vatrapu, R., Medina, R., Joseph, S., and Dwyer, N. 2008. Beyond threaded discussion: Representational guidance in asynchronous collaborative learning environments. *Computers & Education 50*, 4, 1103-1127.

[31] Toulmin, S. 1958. *The Uses of Argument*. Cambridge University Press.