How to Harvest Collective Wisdom for Complex Problems:
An Introduction to the MIT Deliberatorium

Mark Klein
MIT Center for Collective Intelligence

Abstract
Humanity now finds itself faced with pressing and highly complex problems – such as climate change, the spread of disease, international and economic insecurity, and so on - that call upon us to bring together large numbers of experts and stakeholders to deliberate collectively on a global scale. Such large-scale deliberations are themselves complex processes, however, with emergent properties that often prevent us from adequately harnessing the community's collective wisdom. Collocated meetings, beside being expensive, are prone to such well-known effects as polarization, power dynamics, and groupthink. Social media (such as email, blogs, wikis, chat rooms, and web forums) enable more concurrent input but still typically generate more heat than light when applied to complex controversial topics. Large-scale argumentation systems represent a promising approach for addressing this important challenge, by virtue of providing a simple systematic structure that radically reduces redundancy and encourages clarity. This paper will describe the efforts we have made to explore this approach, giving an overview of the key underlying concepts, the ways we have translated these concepts into a working system (the MIT Deliberatorium), and the experiences we have had evaluating this system in a range of contexts.

The Challenge
Decision-making in large communities rarely fully harvests the collective wisdom of its members, even for very high-stakes problems where it can make the difference between disaster and success. It can be simply too expensive to bring the key players into one room, and too difficult to manage the interactions of large groups to get the best that the members have to offer. Only one person can talk at a time, loud voices can dominate a discussion, and emergent dynamics can lead such groups to either deadlock without a solution (polarization) or prematurely settle on a solution without sufficiently exploring the space of promising alternatives (groupthink) [1].

In recent years, social media technologies (e.g. email, web forums, chat rooms, blogs, wikis, and idea forums) have emerged that have the potential to address this important gap. Such tools have enabled diverse communities to weigh in on topics they care about at unprecedented scale, in turn leading to remarkably powerful emergent phenomena [2] [1] [3] [4] such as:

• Idea synergy: when users share their creations in a common forum, it can enable a synergistic explosion of creativity as people develop new ideas by combining and extending ideas that have been put out by others.
• *The long tail*: social computing systems enable access to a much greater diversity of ideas than they would otherwise: “small voices” (the tail of the frequency distribution) that would otherwise not be heard can now have significant impact.

• *Many eyes*: social computing efforts can produce remarkably high-quality results by virtue of the fact that there are many eyes continuously checking the shared content for errors and correcting them.

• *Wisdom of the crowds*: large groups of (appropriately independent, motivated and informed) contributors can collectively make better judgments than the individuals that make them up, even experts, because their collective judgment cancels out the biases and gaps that individuals are prone to have on their own.

Social media technologies often create, however, more heat than light when applied to complex, controversial problems:

• *Disorganized content*: Existing social media generally create very disorganized content, so it's time-consuming to find what has been said on any topic of interest. This fosters unsystematic coverage, since users are often unable to quickly identify which areas aren't yet well-covered and need more attention.

• *Low signal-to-noise ratio*. Social media content is notorious for producing highly redundant content, so important points often got lost in the crowd.

• *Quantity rather than Depth*. Social media systems often elicit many relatively small contributions, rather than a smaller number of more deeply-considered ideas, because collaborative refinement is not inherently supported.

• *Polarization*: Users of social media systems often self-assemble into groups that share the same opinions, so they see only a subset of the issues, ideas, and arguments potentially relevant to a problem. People thus tend to take on more extreme, but not more broadly informed, versions of the opinions they already had.

• *Dysfunctional argumentation*: Existing social media systems do not inherently encourage or enforce well-considered argumentation, so postings are often bias- rather than evidence- or logic-based.

Social media technologies thus often capture only a fraction of the collective wisdom of a community, and enormous effort is typically required to "harvest" this wisdom to inform better, more broadly-supported decisions. Intel, to give a typical example, ran a web forum on organizational health that elicited a total of 1000 posts from 300 participants. A post-discussion analysis team invested over 160 person-hours to create a useful summary of these contributions (at 10 minutes a post, probably longer than it took to write many of the posts in the first place). The team found that there was lots of redundancy, little genuine debate, and few actionable
ideas, so that in the end many of the ideas they reported came from the analysis team members themselves, rather than the forum\(^1\).

It could be argued that many of these concerns are addressed by *topic-centric* tools such as wikis, *where* each discussion topic is captured in it’s own unique, collaboratively-authored, article. But wikis are deeply challenged by complex and controversial topics [5] [6]. They capture, by their nature, the “least-common-denominator” *consensus* between many authors (any non-consensus element presumably being edited out by those that do not agree with it). This consensus is often achieved only by dint of wasteful “edit wars” as different authors repeatedly undo each other’s contributions, and the controversial core of deliberations are typically moved to massive talk pages for the article, which are essentially time-centric venues prone to all the limitations we noted above. To give a concrete example, at the time of writing, the wikipedia “global warming” article was 6500 words long, but the discussion area for this article consisted of 64 archives, each ranging from 2,000 to 60,000 words in length. Articles for more controversial topics are often “locked” altogether to prevent sabotage. Indeed, if we look at the effort users allocate to different wikipedia activities, it appears that the site is increasingly being bogged-down in activities that deal with contention over article contents:

![Changing percentage of edits over time in Wikipedia, showing decreasing direct work (article edits) and increasing indirect work (article talk and so on). [5]](image)

\(^1\) Based on personal communication with Catherine Spence, Information Technology Enterprise Architect, Computing Director/Manager at Intel.
Another promising social media technology is *question-centric tools* such as Dell’s Ideastorm.com, the Obama administrations’ Open for Questions web site, solution contest sites such as innocentive.com, and Google’s project10tothe100.com. Such tools are organized around *questions*: a question is posted and the community is asked to contribute ideas for how to answer that question. Such sites can elicit huge levels of activity — the Obama site for example elicited 70,000 ideas and 4 million votes in three weeks — but they are prone to several serious shortcomings. One is *redundancy*: in all of these sites, many of the ideas represent minor variations of each other. When the volume of posts is large, important ideas that happen to appear relatively few times can be easily overlooked. Pruning such lists to consolidate equivalent posts is, moreover, a massive undertaking. In Google’s case, for example, the company had to recruit 3,000 employees to filter and consolidate the 150,000 ideas they received in a process that put them 9 months behind their original schedule. One could argue that asking the user community to rate submissions can allow the best ideas to rise to the top, but unfortunately ratings systems are prone to dysfunctional ranking dynamics in such contexts. When people are asked to rate a very long list of items, one can expect that the system will quickly “lock” into a fairly static, and arbitrary, ranking. People are more likely to vote for ideas that already have high ratings and, if there are thousands of ideas, people will in all likelihood stop looking after the first few. So the first few winners take all, even if they are inferior to many other ideas in the list. Researchers observed this property when looking at music markets [7]: when people could see each other’s rankings, much of the ranking results were essentially random. A final issue is *depth vs breadth*. Idea aggregation sites tend to elicit many fairly simple ideas. The ideas generated by the google project, for example, (e.g. make government more transparent, help social entrepreneurs, support public transport, create user-generated news services) were in large part not novel and light on detail. The Better World Campaign used an idea sharing system to come up with a proposal for Obama’s first actions upon entering office. After a year of debate, over 5,000 idea submissions, and close to one hundred thousand votes, the result was the proposal that Obama plant an organic garden at the White House. Surely that massive amount of effort could have been used to compose a smaller number of more deeply-considered ideas, but idea aggregation sites provide no support for this, because people can not collaboratively refine submitted ideas.

To summarize: social media technologies enable unprecedented levels of community *input* but often capture only a fraction of the community’s collective *wisdom*, especially with complex and controversial topics, and enormous effort is typically required to “harvest” this wisdom from the large and messy corpuses that result from using these tools.

**A Solution: Large-Scale Argumentation**

Large-scale argumentation represents a promising approach to transcending these limitations and realizing the enormous potential social media has for enabling better decision-making. The approach is simple. Members of a community make their contributions in the form of a
deliberation map, a tree-structured network of posts each representing a single unique issue (question to be answered), idea (possible answer for a question), or argument (pro or con for an idea or other argument):

A screenshot from the Deliberatorium, the large-scale argumentation system developed by the author. Each line in the left pane represents a single issue, idea, or argument. A post’s contents can be viewed, edited, discussed, and rated in the right pane.

Deliberation maps have many advantages over other social media. Every unique point appears just once, radically improving the signal-to-noise ratio. All posts appear under the posts they logically refer to, so all content on a given question is co-located, making it easy to find what has and has not been said on any topic, fostering more complete coverage, and counteracting polarization by putting competing ideas and arguments right next to each other. Careful critical thinking is encouraged, because users are implicitly encouraged to express the evidence and logic favoring the ideas they prefer, and the community can rate each separate element of an argument individually. A key element of the Deliberatorium is the “live and let live” rule: if one disagrees with an idea or argument, the user should not change that post to undermine it, but should rather create new posts that present the strongest ideas and arguments they can muster, so all contributions can compete on an even basis. Users, finally, can collaboratively refine
proposed solution ideas. One user can, for example, propose an idea, a second raise an issue concerning how some aspect of that idea can be implemented, and a third propose possible resolutions for that issue. The value of an argument map can extend far beyond the problem instance it was initially generated for: it represents an entire design space of possible solutions that can be readily harvested, refined and re-combined when similar problems arise at other times and places.

The Deliberatorum utilizes a simple process to ensure that the deliberation map built by the community is well-structured:

![The map-building process used by the Deliberatorum.](image)

Every author submission should represent a single issue, idea, or argument, should not replicate a point that has already been made elsewhere in the argument map, and should be attached to the appropriate part of the map. Posts should only be edited to strengthen them (the “live and let live” rule): if one disagrees with an idea or argument, the user should create new posts that present their alternative ideas or counter-arguments. Moderators, a common feature of many social media systems, help ensure that these guidelines are followed. New posts can, at first, only be viewed by moderators and other authors. When a moderator verifies that the posts follow the deliberation map guidelines, they become certified and can be viewed, edited, commented, and rated by the full community. If a post doesn’t yet meet the guidelines, the moderator leaves comments explaining what needs to be done to fix them. In all this, moderators play a relatively modest “honest broker” role: their job is not to evaluate the merits of a post, but simply to help authors ensure that the content is structured in a way that maximizes its utility to the community at large.
We estimate, based on our experience to date, that there needs to be about 1 moderator for every 20 active authors, to ensure that posts are checked and certified in a timely fashion without undue burden on each moderator. This figure is well within the bounds of the percentage of “power users” (also known as “meta-contributors”) that are a common feature of many successful peer-production systems, including Wikipedia, Slashdot, and so on. Experienced authors, with a track record of successful argument map creation, can be selected to join the moderator pool. There is already, in addition, a substantial world-wide community of people with argument mapping skills. One organization alone (cognexus.org) has trained and certified hundreds of people on argument mapping techniques. Argument mapping is, in addition, a natural skill for lawyers, philosophers, mathematicians, library scientists, debaters, and others who frequently create arguments or proofs. Such individuals may be inspired by the opportunity to contribute their skills to deliberations about complex critical challenges, even if they do not have substantial content expertise in that area.

While this has not happened in our evaluations to date, it is of course possible that moderators may undercut the deliberations by imposing their own biases when deciding which posts to certify. There are many ways to address this concern. For example, moderators can double-check each others’ decisions, especially when authors complain that a post of theirs was inappropriately left uncertified. We can also provide a moderator rating system so that the user community can identify consistently poor moderators, as well as a voting process for removing moderator privileges for such individuals. Such a “meta-moderation” model has worked successfully with such systems as Wikipedia and Slashdot.

This map-building process is supported by such software capabilities as open editing (any user can check and improve posts), watchlists (which automatically notify users of changes to posts they have registered interest in) and version histories (to allow users to roll-back an post to a previous version if it has been “damaged” by an edit).

**How Well Does Large-Scale Argumentation Work?**

*The Carbon Offsetting Thought Experiment:* Our first effort to evaluate the viability of a large-scale argumentation approach was a thought experiment that involved translating a web forum on carbon offsetting (hosted on planet.com on May 2008) into a deliberation map. The original 13 page discussion, filled with the digressions and repetitions that typify web forums, translated into the following map:
Clearly, in this instance, the eight item deliberation map represented a much more usable rendition than the original 13 page forum. This was a startling initial illustration of the potential of deliberation maps for harvesting a community’s collective knowledge in a way that is qualitatively more useful than conventional media.

*The Naples Evaluation:* Our first major evaluation of the Deliberatorium system itself was at the University of Naples, where 220 masters students in the information engineering program were asked to weigh in, over a period of three weeks, on what use Italy should make of bio-fuels. This study helped address key uncertainties about the viability of a large-scale argumentation approach. We were concerned that the additional effort required to add to a deliberation map might quell user participation, and that most users would be unable to properly structure their posts, placing an insupportable burden on the moderators. Neither of these concerns proved to be the case. We observed a very high level of user participation. Remarkably, the Deliberatorium was active almost continuously, except for a daily hiatus between roughly 3 and 6 am:
Growth in number of posts over the three weeks of the Naples evaluation.

About 180 out of 220 users contributed at least a few posts, and the most active contributed 40 or more posts each:

The number of contributions per user in the Naples evaluation.

In two weeks the students posted nearly 3000 issues ideas and arguments (of which roughly 1900 were eventually certified) in addition to over 2000 comments:
<table>
<thead>
<tr>
<th>Type of Post</th>
<th># of Posts</th>
<th># of Certified Posts</th>
<th>% (certified)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Issue</td>
<td>242</td>
<td>89</td>
<td>5%</td>
</tr>
<tr>
<td>Idea</td>
<td>962</td>
<td>452</td>
<td>24%</td>
</tr>
<tr>
<td>Pro</td>
<td>1488</td>
<td>1045</td>
<td>55%</td>
</tr>
<tr>
<td>Con</td>
<td>402</td>
<td>325</td>
<td>17%</td>
</tr>
<tr>
<td>Comment</td>
<td>2009</td>
<td>n/a</td>
<td>n/a</td>
</tr>
<tr>
<td>Grand total</td>
<td>5003</td>
<td>1911</td>
<td>100%</td>
</tr>
</tbody>
</table>

The system supported a lively debate amongst the students: about 70% of the arguments, for example, addressed posts created by other authors.

All this was achieved with the part-time support of just two moderators. This was possible because most posts (about 60%) were structured correctly by the authors when they created them (and thus could be certified without changes), and a further 30% only required one input from the moderator before certification:

![Moderator inputs needed before certification](image)

*Number of moderator inputs needed to certify a post, Naples study.*

The average number of moderator inputs needed to certify a post decreased significantly, about 35%, over the course of the evaluation, suggesting that the community became increasingly skilled with the argument mapping approach:
Change in number of moderator inputs needed before a post was certified. The reduction of 35% is statistically significant at $p < 10^{-9}$.

In the end, the students created a map that was judged by content experts to represent a remarkably comprehensive and well-organized review of the key issues and options around bio-fuel adoption, exploring everything from technology and policy issues to environmental, economic and socio-political impacts:

**Deliberation Map**

- How to reduce the barriers to the spread of biocidal products...
  - reducing the impact on the environment and health...
  - through appropriate subsidies from the state apre
  - increase the availability of local raw materials...
  - By introducing appropriate sanctions
  - reducing conflicts and local self-interest
  - Weakening of the opposition of the multinationals...
  - reducing the "competition for land" and increasing...
- What future for biofuels in Italy?
  - The future of green energy committee is already in Italy...
  - How to enhance the production of bi...
  - How can Italy 'supply and use'...
  - To what extent and how the production of biochar...
  - To what extent will spread in Italy 'use'...
  - What effects the development of biofuels try...

A portion of the Naples bio-fuels deliberation map (translated from Italian)
We were hard-pressed to imagine any other approach that would allow over 200 authors to write what was in effect a substantial book on a complex subject, in a couple weeks, with no one in charge.

We have since developed a better theoretical understanding of why we were able to achieve such a substantial level of use in a large-scale argumentation system. Active participation in social media systems tends to occur when the benefits, to the individual contributors, substantially exceed the costs [8]. With small numbers of participants, clearly an informal approach, based for example on phone or email or web forums, minimizes participation costs and is capable of producing good deliberation outcomes. But this picture, we suggest, changes as the scale of the discussion grows.

It has been found [9] [10] [11] [12] that users contribute to social media systems predominantly for two reasons: (1) to find their tribe (i.e. get connected with people who share their interests) and (2) to become a hero (have a substantive positive impact on a community they care about). How does this play out in a deliberation context? Let us make the reasonable assumption that points (individual issues ideas, and arguments) have an uneven distribution in the user population, i.e. some points are known to most people, some to only a few. We can thus expect widely known points to be submitted frequently from multiple sources, and more “out-of-the-box” (but potentially valuable points) to arise less often. It seems clear that the number of unique points contributed to the deliberation will grow much more slowly than the number of participants. Our simulations suggest that this growth is roughly a logarithmic function of the community size, implying a roughly linear growth in redundancy.
Simulation results for number of redundant posts in a deliberation.

The larger the user community, therefore, the more potential redundancy there is, and thus the more value argument mapping offers in terms of improving the signal to noise ratio. There is widespread disaffection with the low signal-to-noise ratio of current social media tools. We can thus expect that, as the scale of the discussion grows, users will increasingly recognize the opportunity to “become a hero” by contributing something (i.e. creating a value-rich deliberation map) that is highly valued by the community. Argument mapping also increases user’s chances of “finding their tribe”. While contributing to unstructured discussions is easier, the high volume and redundancy of such discussions means that most posts will probably be overlooked by most readers. In an argument map, by contrast, if you have a unique point to make, it has a much greater chance of being seen. We can thus expect the benefits of argument mapping, to a contributor, will increase rapidly with the size of the user community.

We can also expect, moreover, that the costs of participation for a contributor will grow only slowly as the community scales. Contributing to a deliberation map incurs two main costs:
1) *unbundling* the contribution into its constituent issues, ideas, and arguments
2) *locating* the proper place for these elements in the map

The cost of unbundling a contribution is independent, of course, of the size of the map. The cost of *locating* a contribution should increase with the size of the map, but only slowly. Remember that a deliberation map is structured like a tree. To find the right place to put a post in a tree, you just have to pick the right top-level branch to place it, the right sub-branch under that, and so on, until you reach the place where it belongs. If the average branching factor (number of sub-branches per branch) of a tree is $N$, then the average number of steps needed to locate a post is the $N$th logarithm of the tree size. The overall picture is thus the following:

![Cost vs Benefits](image)

*Qualitative model of cost vs benefit tradeoffs for contributing to an argument map as a function of the size of the active user community.*

The *benefits* of adding to an argument map grow as the community scales, in proportion to how much the signal-to-noise ratio is improved, roughly linearly. The *costs* of adding to an argument map grow only logarithmically with scale. At some point, we can expect, the benefits to
individual authors will greatly exceed the costs, thereby providing compelling incentives for participation.

What about moderation? Who will do it? Will we be able to find enough moderators? Power users, with a track record of successful argument map creation, can be recruited to join the moderator pool. We estimate that there needs to be about 1 moderator for every 20 active authors, to ensure that posts are checked and certified in a timely way without undue burden on each moderator. This figure is well within the bounds of the number of “power users” that emerge naturally in existing web 2.0 systems. Scale helps us here too. The number of potential moderators should grow with the size of the community, but the number of posts to certify should increase a lot slower than that, as people increasingly find that their points are already captured in the map. So the per-moderator burden should decrease as the community scales. A final key point is that an organization will probably prefer to make a relatively modest investment in up-front moderation in order to avoid the much larger costs of harvesting a discussion after it has occurred.

Technological and process refinements can further improve this picture by reducing authoring and moderation costs. One option, for example, is “wisdom of the crowds” moderation, wherein the full user community, performing simplified moderation “micro-tasks”, substitutes for a relatively small cadre of expert moderators. There is good reason to believe that such an approach can work. It has been shown many times that large numbers of people with even modest skills can, under some circumstances, in the aggregate perform judgment tasks better than experts [Surowiecki, 2005 #4666]. It has also been shown, by such systems as Amazon’s Mechanical Turk, that large number of people are willing to perform such micro-tasks cheaply, or even for free if they believe in the project.

*The Intel Evaluation:* Our next key evaluation assessed how a large-scale argumentation approach will fare in the “wild”, outside of the confines of a classroom environment. We helped Intel Corporation conduct a deliberation on how “open computing” (i.e. where users are given greater access to computing tools and data) should be used in the company. The discussion was advertised broadly, but contributions were purely voluntary. A single moderator was able to support the discussion with minimal effort. The end result (see below) was that Intel received a substantive and well-organized overview of important issues in this space from 73 contributors, including many from outside the company, at almost zero cost.
Conclusions: In addition to the evaluations described above, we also conducted evaluations with the US Bureau of Land Management, the University of Zurich, and HMC Inc, among others. Our conclusions from this work are the following. While the Deliberatorium is of course not the be-all and end-all of collective decision-making – a community may decide for example to ignore the top-rated ideas and arguments in the map in favor of their own initial intuitions - it does appear that it enables you to get a more complete picture of how to solve complex problems, with far less effort, than conventional social media.

Future Work
While the work to date is promising, there remain many important areas for future development in the realm of using large-scale argumentation to support effective large-scale deliberation. We
will highlight two such areas that are being investigated actively by the author and his collaborators:

Attention mediation: Even a moderately large community can quickly generate large deliberation maps when deliberating about complex topics. How can we help users identify the portions of the map that can best benefit from their contributions, in maps that cover hundreds of topics? How can the stakeholders for such deliberations assess whether the deliberations are progressing well, whether some intervention is needed to help the deliberations work more effectively, and when the results are mature and ready to “harvest”? Can we foster, for large-scale deliberations, the understanding that participants in small-scale discussions typically have about where the discussion has gone, what remains to be addressed, where the trouble spots are, and where they can best contribute? Without this kind of big picture, we run the risk of severely under-utilizing the collective intelligence potentially accessible by large-scale social media. We are currently investigating how automatically-derived deliberation metrics can enable better attention allocation decisions. This has proved to be a fertile area; deliberation maps offer a unique opportunity to define novel and useful metrics because they capture a significant degree of semantics (post types and links) in addition to the keyword frequency and social network information available with standard social computing tools. We can, for example, derive metrics that help us identify how mature a deliberation is (deliberations tend to move from issue identification to idea brainstorming to idea evaluation), what the controversial issues are (there are many opposing ideas and arguments, with deep argument chains), whether some elements of the discussion evince a high level of bias (e.g. people give high ratings to ideas whose underlying arguments are poorly rated – see [13]), whether some elements of the discussion have become highly balkanized (i.e. there are distinct groups of participants that talk mainly to each other and have opposed positions on a given issue), whether there is excessive groupthink (community attention is focused on a small subset of the options for an issue before the other options have been much explored), and so on. For more detail, see [14].

Moderation: A key challenge with the large-scale argumentation approach is the need for moderators to help ensure that contributors produce a well-structured deliberation map despite the fact that most authors will not be initially familiar with the formalism. While it is reasonable to expect that super-users with strong moderation skills will emerge from large long-lived deliberation communities, just as they have from other social media communities, this does not help if we have shorter-lived deliberation engagements and, in any case, wider adoption of a large-scale argumentation approach will be facilitated if we can reduce the moderator burden as much as possible. We are exploring, for this purpose, a combination of search tools (to help authors and moderators quickly find where to locate their posts) as well as “wisdom of the crowds” moderation (wherein large numbers of people, performing simplified micro-tasks, substitute for a relatively small cadre of expert moderators). There is good reason to believe that such an approach can work. It has been shown many times that large numbers of people with even modest skills can, under some circumstances, in the aggregate perform judgment tasks better than experts [3]. It has also been shown, by such systems as Amazon’s Mechanical Turk, that large number of people are willing to perform such micro-tasks cheaply, or even for free if they believe in the project.
References