Do You See What I See? Perception and Navigation in Online Deliberation

W. Ben Towne

CMU-ISR-17-101

April 2017

Institute for Software Research
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213 USA

Thesis Committee
James D. Herbsleb (Chair)
Carolyn P. Rosé
Daniel B. Neill (Heinz)
Thomas W. Malone (Massachusetts Institute of Technology)

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Societal Computing

This research was sponsored by the National Science Foundation (NSF) under grant numbers IIS-1302522 and IIS-1111750, as well as by NSF's XSEDE (eXtreme Science and Engineering Discovery Environment) under Contract No. FA8721-05-C-0003 with Carnegie Mellon University for the operation of the Software Engineering Institute (SEI), a federally funded research and development center sponsored by the Department of Defense, by a LENS grant from the SEI, and funds from the Institute for Software Research at Carnegie Mellon University.

The views, findings, and conclusions contained in this document are those of the author and should not be interpreted as representing the official policies, either expressed or implied, of the National Science Foundation or the U.S. Government.

Keywords: Computer science, online deliberation, collaboration at scale, collective intelligence, crowdsourcing, perception, perceived quality, empirical methods, navigation, navigability, text corpora, social influence, experiments, comments, distributed evaluation, creative work, peer production.

Abstract

Some of the most pressing challenges facing humanity today, such as how to respond to climate change or govern the internet, are too complex for any individual or small group to completely understand by themselves, but a lot of people each know a piece of the problem or solution. This is somewhat like a billion-piece jigsaw puzzle where somebody threw away the box and mailed each piece to a different person. Attempts are now being made to build platforms where people can bring their pieces and assemble them into solutions.

This thesis examines such platforms, how people use and perceive them and their content, and how certain design decisions such as the exposure of discussion behind collaboratively produced content can affect those perceptions. Through a set of studies ranging from qualitative interviews to controlled experiments with hundreds or thousands of participants, this thesis adds to our understanding of how humans use and perceive content on such platforms, including when primary content is presented alongside related content or discussion, so that we can better understand how to design these systems to better achieve their users' intended goals.

The research integrates insights from computer science and social psychology with latent variable modeling techniques in order to increase our understanding of what people are trying to accomplish on an example platform designed to support large-scale collaboration around a complex issue, and experimentally explores how people perceive the content they find on such sites. Data and projects used in this thesis come from a diversity of sources including Wikipedia, the President's SAVE award ideation contest (facilitated through the IdeaScale ideation platform), and the MIT Climate CoLab.

Table of Contents

Abstract 3

Thesis Committee 7

Acknowledgements 8

Summarizing Introduction 11

- Motivation 11
- Work 12

Study 0: Background 17

- 0.1: Background: Summary 17
- 0.2: Background: Online Deliberation and Technologies to Support it 18
- 0.3: Background: Explicit Formal Structuring and Linking Of Text in Online Deliberation Settings 24
- 0.4: Background: Navigation through Textual Corpora 36
- 0.5: Background: Overview 50

Study 1: Interview Study of Platform Participants 51

- 1.1: Study 1: Introduction 51
- 1.2: Study 1: Domain 54
- 1.3: Study 1: Position in Framework 55
- 1.4: Study 1: Method 56
- 1.5: Study 1: Results 63
- 1.6: Study 1: Discussion 72

Study 2: Evaluating a State-of-the-Art Topic-Modeling Based Similarity Measure 75

- 2.1: Study 2: Chapter Summary 75
- 2.2: Study 2: Introduction 76
- 2.3: Study 2: Literature on Evaluating Similarity Measures 77
- 2.4: Study 2: Experiments: Overview 81
- 2.5: Study 2: Models And Materials Selection 82
- 2.6: Study 2: Experiment 1 94
- 2.7: Study 2: Experiment 1b 98
- 2.8: Study 2: Experiment 2 101
- 2.9: Study 2: Experiment 2b 104
- 2.10: Study 2: Experiment 3 107
- 2.11: Study 2: Limitations 111
- 2.12: Study 2: Conclusions 112

Study 3: Evaluating the Effects of Exposing Details of Deliberative Process on Readers 115

- 3.1: Study 3: Chapter Summary 115
- 3.2: Study 3: Introduction 115
- 3.3: Study 3: Conflict Resolution Strategies 117
- 3.4: Study 3: Wikipedia Talk Pages 118
- 3.5: Study 3: Judging the Credibility of Information Online 119
- 3.6: Study 3: Experiments 124
- 3.7: Study 3: Results 129
- 3.8: Study 3: Conclusions and Discussion 134

Study 4: Conflict in Comments: Learning and the Limits of Carry-Over 139

4.1: Study 4: Chapter Summary 139
4.2: Study 4: Introduction 139
4.3: Study 4: Method Overview 141
4.4: Study 4: Hypotheses 141
4.5: Study 4: Materials 143
4.6: Study 4: Manipulation Check on Comments 151
4.7: Study 4: Main Experiment Results 154
4.8: Study 4: Conclusions 160
5. Conclusions 163
5.1: Brief Summary of Study Findings 163
5.2: Discussion of Contributions 163
5.3: Future Work 165
5.4: Broader Implications 169
Glossary 171
References 173

Readers of electronic versions of this thesis can click (or CTRL+click) to jump to a desired section. Within-document hyperlinks should also work for most cross-references (e.g. sections, figures, etc.) Reader software with an outline/navigation pane can also be used for a more detailed Table of Contents.

Thesis Committee

James D. Herbsleb (chair)
Director, Societal Computing PhD Program in ISR
CMU School of Computer Science

Carolyn P. Rosé
Language Technologies Institute & HCII
CMU School of Computer Science

Daniel B. Neill
Director, Joint PhD Program in Machine Learning & Public Policy
CMU H.J. Heinz III College

Thomas W. Malone
Director, MIT Center for Collective Intelligence
MIT Sloan School of Management

Acknowledgements

It is perhaps appropriate that the preparation for and of a thesis about platforms for large-scale online collaboration has benefited from fruitful collaboration with so many others. This work has been improved by insights and support from many people. Although the singular first person is used in most of the document that follows, there were others who made significant intellectual contributions to it.

First, thanks to my thesis committee, without whose guidance and support this thesis would not have been possible: Jim Herbsleb, Tom Malone, Daniel Neill, and Carolyn Rosé. Thanks also to additional co-authors on papers written during my time at CMU even if not directly part of this thesis, including (again in alphabetical order) Christian Adriano, Claudia Müller-Birn, Patrick de Boer, Eric Chiquillo, Laura Dabbish, Jana Diesner, Laur Fisher, Yue Han, Andre van der Hoek, Peter Kinnaird, Aniket 'Niki' Kittur, Thomas LaToza, Robert Laubacher, Arturo Di Lecce, Jeffrey Nickerson, Pinar Ozturk, and Fabio Ricci.

Thanks to other formal and informal members of the Herbsleb Research Group including Chris Bogart, Marcelo Cataldo, Chalalai "Jib" Chaihirunkarn, Anna Filippova, Pranav Gupta, James Howison, Arun Kalyanasundaram, Jon Kush, Erik Trainer, Jason Tsay, Marat Valiev, Bogdan Vasilescu, Patrick Wagstrom, and Evelyn Zhang. Thanks to members of Carolyn Rosé's TELEDIA lab, including Hyeju Jang, Keith Maki, Oliver Ferschke, Michael Miller, and Miaomiao Wen. Thanks to additional colleagues in the Human-Computer Interaction Institute including Jeff Bigham, Joel Chan, Steven Dow, Walter Lasecki, and Jenn Marlow. Thanks to Gary Olson, Nancy Taubenslag, and other contributors to the Climate CoLab project. Thanks to the other members of the SCALE collaboration across Carnegie Mellon University, University of Nebraska at Lincoln, and the University of California at Irvine. It has been helpful to bounce around ideas about work in progress and benefit from your perspectives on not only the specific studies referred to in this thesis but in-progress research more broadly.

Thanks to Jesse Dunietz, Daniel Gingerich, Adona Iosif, Kelly Matula, Ardon Shorr, Junjue Wang, Julian Whitman, and others at Public Communications for Researchers for helping me be able to

communicate this and other work more clearly than I might otherwise be able to. Thanks to Vishal Dwivedi, Hanan Hibshi, Michael Maass, Ivan Ruchkin, and Jason Tsay for extensive comments on drafts in our thesis-writing groups. Thanks also to many anonymous reviewers for extensive comments on other write-ups of the same studies in the peer review process, which helped improve the work.

Thanks to Mark Whiting and members of Daemo (Gaikwad et al., 2015) for help with piloting some Amazon Mechanical Turk tasks. Thanks to the many Turkers who participated in either the main experiment or manipulation checks and the Climate CoLab members who contributed their time and perspectives during interviews, group sessions, and informal chats. Thanks to SAVE award contributors and facilitators both in the federal government and at IdeaScale.

Thanks to Mark Klein at MIT as well as M. Bernardine Dias and Manuela Veloso at CMU for supervision of exploratory review in my first semester at CMU, which helped produce some of the Background and motivation material in this thesis as well as the relevant journal article.

The staff of Carnegie Mellon's Institute for Software Research have also been instrumental in supporting this work and helping it go more smoothly or easily than it otherwise might have. Thanks to administrative support staff Sharon Blazevich, Catherine Copetas, Nick Frollini, Connie Herold, Helen Higgins, Janice Kusmierek, Victoria Poprocky, Josh Quicksall, and Monika De Reno. Thanks to ISR IT support directors Chris Dalansky and Tom Pope and technicians Emanuel Bowes and Ryan Johnson as well as the SCS and CMU-wide IT support teams. Thanks to Eric Rosé for programming experimental infrastructure supporting many of the experiments that went into this thesis.

Thanks to Eden Fisher, Karen Fleischman, & Jimmy Williams at the Engineering & Technology Innovation Management program, as well as Professor of the Practice Deborah Stine, for allowing me to take part in your courses examining practices that promote innovation in large organizations and settings today. Thanks also to professor Mark Roosevelt, Debra Tekavec and the Students for Science &

Technology Policy, as well as Sally Stadelman of the Mayor's Civic Leadership Academy and others in the City of Pittsburgh (especially the Department of Innovation & Performance) for additional very enlightening perspectives on policy evaluation. Some of these perspectives led directly to the motivations and even specific questions used in the studies below.

Thanks to Professor Robert Cavalier in CMU's Philosophy department for support and experienced guidance in pioneering online deliberation and continuing to seek effective ways of facilitating in-person conversations with mid-sized groups about sometimes complicated issues. Thanks to Sandy Heierbacher and many others at the National Coalition for Dialogue & Deliberation for advancing the field and building a national network of practitioners and platform innovators as well as a public more actively prepared to engage with platforms and processes like those this thesis helps advance.

Thanks to the CMU administration for supporting this work, especially VP of Student Affairs Gina Casalegno, Assistant Vice Provost for Graduate Education Suzie Laurich-McIntyre, Assistant Director for Student Affairs M. Shernell Smith, and Assistant Dean of Student Affairs Liz Vaughan, for direct support.

Thanks to Congress and the American public for financially supporting this and other developments in science, even when scientific advances do not always support particular political positions. Finishing this thesis around the start of the Trump administration has served as a reminder this cannot always be taken for granted (Glass, 2016). Thanks especially to the National Science Foundation who funded this work through grants IIS-1302522, IIS-1111750, and Contract No. FA8721-05-C-0003, as well as the Software Engineering Institute at Carnegie Mellon University.

Finally, thanks much to friends and family who have supported me throughout this whole process. The completion of this work would not have been possible without it, and I am very grateful. Thank you.

Summarizing Introduction

Motivation

We live in an increasingly diverse yet highly connected society, which presents both challenges and opportunities (Shlain, 2011). On the positive side, we have the possibility to bring together a broad assortment of experience, skills, and expertise, which offers the potential for attacking big problems in diverse and creative ways. Participation of many individuals is required in order to take advantage of this tremendous resource. In order for the value in the diversity (Page, 2008) to produce creative solutions, or for problems that exceed the complexity limits of any individual or small group to be solved, many differently oriented people will need to work together. Different people may see or understand different parts of a problem. Even within existing organizations, "no one at the top of large organizations can ever know enough to see and understand all the needs and potentials for change." (Malone, 2004, p. 165). It is argued that humans' ability to cooperate both flexibly and in very large numbers is the differentiating factor that makes us the most successful species on Earth (Harari, 2016, sec. 2:06-4:06). We need to coordinate up to billions of individuals (Harari, 2016) but still have some work to do before we can coordinate flexibly at a large enough scale to adequately address our greatest remaining challenges.

For many large, distributed problems, there exists a large, distributed group of people who each understand and are willing to contribute part of the problem perception or solution. People are looking to more actively engage with the world around them and contribute to these efforts, and we're just starting to understand how to tap into that resource and aggregate these little pieces together to create something useful. We now have at least most of the building blocks needed to build a platform where people can bring their proverbial piece of a puzzle and see where it fits in and which other pieces that connect most closely to it, in the informed hopes that this will help us build up a sense of the bigger picture and be able to address some of these issues, even if a full understanding remains beyond the capacity of any individual. Ideation tools and other open platforms now help present challenges and gather a lot of ideas (often presented as text) that people think will be helpful in solving these problems.

However, even a warehouse full of pieces isn't all that useful until they can be connected together, bit by bit, so that a whole large group of people in parallel can help solve the puzzle. In current systems, the pieces are often not very organized, and it takes a long time to go through the ideas.

Work

The need to better support collaboration in large diverse communities is well known but still unsolved. This thesis explores, develops, and tests hypotheses that increase scientific understanding that could help address a focal problem within this larger space, specifically the problem of supporting meaningful idea exchange within diverse communities and groups in the form of online deliberation. Even more specifically, this thesis examines how participant perceptions (such as evaluations of content quality) are affected by algorithmic and interface design choices for presenting content (such as supporting dynamic connections based on topical similarity and the relative visibility of discussions behind content) on platforms for large-scale ideation and collaboration.

While complex real-world policy problems become increasingly critical to address, they are now often paired with increasing size and complexity of relevant data, including a large volume of free-text data, and a drive toward new and transformative participatory technologies. The work in this thesis advances the science behind online deliberation, collective intelligence, and large-scale online collaboration towards public good, better enabling the design of successful systems that may be able to help large, distributed groups leverage diverse knowledge and perspectives to address complex issues.

This thesis is organized in six chapters (including conclusion) as follows. First, a background section summarizes lessons from the literature and previous work. This section draws on a literature review published in the *Journal of Information Technology & Politics* (Towne & Herbsleb, 2012) as well as material that introduced later studies in their respective papers and additional relevant background tied in since those publications.

A sequence of four studies is then described, starting with the most qualitative work and proceeding through more detailed randomized controlled experimental studies. (As the terms are used here, a

given “Study” may contain multiple controlled “Experiments.”) Following the literature review of various platforms for online deliberation and tools for supporting some related tasks, Study 1 is a qualitative interview-based study which deepens our understanding of the variety of goals, motivations, strategies, and perceptions diverse participants may bring to a platform for large-scale collaboration around a complex issue like climate change, and explores how some existing solutions may or may not help support these objectives. Some results from this work were presented at the *Conference on Collective Intelligence* (Towne, Rosé, & Herbsleb, 2016b). One reported obstacle to greater impact is the challenge of finding people and proposals, for purposes that vary depending on the participant’s role and goal (e.g. finding collaborators or projects to contribute to, identifying best proposals or projects to invest in or support off-platform, etc.). Issues of perceived navigability can make a big difference, even to how motivated authors feel to contribute content.

One strategy that may be appropriate to address those navigability-based challenges and increase the effectiveness of participants on such a platform is adding dynamic cross-links based on similarity of topics discussed in each piece of crowd-contributed content (such as an idea proposed in response to a specific challenge). Algorithms exist to find similarity based on modeling the topics each document contains. Study 2 presents a series of experiments comparing one common algorithmic measure of similarity (LDA-based cosine similarity) against human perceptions of similarity, making novel methodological contributions. This study also goes into greater detail on identifying areas where the two disagree and identifies some of the specific limits such algorithms have in these applications, such as when humans identify similarity by comparing idea quality or understandability, which may be a practical limit of bag-of-words models more broadly. Another written report of this study was published in *Transactions on Intelligent Systems and Technology* (Towne, Rosé, & Herbsleb, 2016a).

Altering a user interface to expose related material alongside whatever a viewer may be looking at may be helpful in addressing some of the challenges identified in Study 1, with qualifiers derived from Study 2, but depending on the *type* of that related material, it can also change the way participants perceive the primary content they are viewing. In addition to dynamically increasing exposure of

topically related content that is the *same type* as whatever content a viewer might be looking at in a given moment on a particular platform, the *discussion behind* that content may also be a relevant source of information affecting how people perceive and interpret the content, though in many current systems the discussion is stored and presented on a separate page away from the most closely related collaboratively produced artifacts. Showing the discussion about a particular piece of content is one special case of showing material that is topically related to that content.

Study 3 experimentally uncovers some details about how increasing visibility of that related discussion changes the way people perceive the resulting content. It presents readers with a section of a controversial but high-quality Wikipedia article, alongside one of several kinds of discussion about that content, or no discussion at all, and asks those readers to rate the article quality. This work finds that the presentation of any discussion can cause readers to rate the article as lower quality than readers who saw no discussion, a result surprising to even those readers. The strength and details of this effect also depend on whether the discussion contained conflict and if so, how that conflict was resolved. Another written report of this work was presented at the *ACM Conference on Computer-Supported Cooperative Work and Social Computing* (Towne, Kittur, Kinnaird, & Herbsleb, 2013).

Study 4 tests these interesting findings in greater detail, using a similar methodology extended to see how far those effects might carry, and testing materials from a different context. It tests to see if comment-caused effects on how people perceived the commented-on content carry over to other proposals by the same author, about similar topics, or simply found on the same platform. Materials used in this study are framed (accurately) as *proposals* for how to address some element of climate change, in contrast with Study 3's use of artifacts mentally modeled by many as a reference work. This study finds that while the type of comment presented still affects how people perceive the content quality, the effects in this setting are more consistent with participant expectations and the mere presence of non-conflict comments is not sufficient to lower perceptions of content quality. It further finds that the comment-caused rating differences do not carry over to a second proposal evaluated by the same reader immediately after the first even if the second proposal is by the same author or about a

similar mix of topics, and that participants report learning more from constructive conflict than other types of comments. Another report about this work has been accepted for publication at the *ACM Conference on Human Factors in Computing Systems (CHI)* (Towne, Rosé, & Herbsleb, 2017).

This thesis contributes to our understanding of the diversity of roles, objectives, and perceptions one might find on platforms for large-scale collaboration around complex issues. It exposes some of the challenges currently experienced by users of such platforms today, raising several questions which can be addressed by the field in future work. It deepens our understanding of the efficacy of some particular algorithmic solutions to the challenge of navigability, helps us better understand some of the fundamental limits of that class of algorithms, and makes methodological contributions that allow us to measure the efficacy of proposed alternatives over the state of the art. It also contributes an experimentally derived understanding of how presenting discussion topically related to peer-produced content affects readers' perceptions of that content. It further presents evidence about the limits of these effects, including experimental results showing how the effects may differ when readers begin with different pre-existing perceptions of the content source or type, as well as when discussion participants use different strategies to resolve conflicts related to the content.

The results of this work raise questions which may be addressed by additional work in the field going forward. How can we best support the particular objectives and challenges of participants in diverse roles on platforms for collaboratively solving complex challenges, like those identified in Study 1? How can we address the limits of algorithmic similarity measures, like those identified in Study 2, and/or utilize the strengths of such measures to complement other strategies for improving navigability in text corpora? What are the underlying psychological explanations for the way the presence of discussion affects the perception of peer-produced content, opposite to the way people think they are affected, as found in the results of Study 3 and differences between studies 3 and 4? Some hypotheses prompted by these studies are discussed in the following thesis document, which can be used as a basis for further advancing work in this field.

Study 0: Background

0.1: Background: Summary

The three background sections below draw highlights and lessons from some of the history of online deliberation, and efforts to organize text contributions in applicable contexts. The first outlines the value of deliberation, especially online, and motivates the need for supporting technologies. It then introduces a framework ("genome") for questions around building collective intelligence systems.

The second part reviews some of the history of how online deliberations have been structured, and how presentation of the discussion has related to presentation of the collaboratively produced artifact/record. It covers Issue-Based Information Systems, their descendants, two-sided deliberation tools, and ideation tools. In order to advance the field of online deliberation, it is appropriate to know these origins and to understand the considerable benefits derived from having this structure, which was often manually created in the absence of a suitable automatic algorithm. These formally structured conversations are also often brought up first when people think about facilitating large-scale online discussions. This section concludes with a discussion about design considerations related to how content is organized and presented.

The third section of background looks at various approaches to navigating through text corpora, including those outside the specific field of online deliberation. It briefly discusses how certain strengths and weaknesses of each affect what users can and do accomplish using that strategy. Subsection 0.4.2 briefly discusses some of the limits of keyword searches, which are presently a dominant method of finding content in large corpora. For reasons discussed there, this thesis does not focus on tweaks to improve keyword search, so this background chapter does not take the large space that would be required for in-depth coverage of the extensive literature on algorithms for or perceptions of keyword search result sets. While the review discusses navigation based on hyperlinks, it does not cover algorithms that rely on extensive pre-existing link structures such as PageRank or bibliometric/citation-

based measures, because such structures are not generally available in online deliberation settings, especially ideation settings where contributors are not reasonably expected to read everything that is already there.

0.2: Background: Online Deliberation and Technologies to Support it

0.2.1: The Value of Deliberation

Deliberation, or rational discourse marshaling evidence and arguments bearing on a decision, brings many potential advantages to the decision-making process (e.g. Mathews, 2014; Sunstein, 2006), as compared to simply applying a decision rule to aggregated expressions of preference. Deliberation is a key component leading to informed and engaged participants who might be able to achieve rich insight into complex problems (Cavalier, 2011). Indeed, online deliberation (Davies, 2011) may be the only way to address the pressing and highly complex problems facing humanity today, such as responding to climate change, governing global commons like the Internet, caring for health, providing resources for a growing global population, and sharing resources productively, because these issues require gathering large numbers of stakeholders to deliberate collectively in a complex space where no single person or committee can even understand all the applicable interdependencies (Klein, 2011a).

When asked to identify one of society's most pressing problems deserving of a major incentive prize, Dr. Shwetak Patel cited the construction of better platforms for large-scale collaboration on complex problems (Zane, 2012). Others go so far as to claim that "Modern-day political participation is dependent on widespread deliberation supported by information and communication technologies, which also offer the potential to revitalize and transform citizen engagement in democracy" (Rose & Sæbø, 2010, p. 228). Online civic engagement is increasing, like online engagement of many forms, but hopes that this would draw a larger and more diverse population into the political process remain

largely unmet (Sciullo, 2013). A 2016 *Science* article asserts that human computation has "huge potential" to address wicked[1] problems, going beyond microtask crowdsourcing to create shared online workspaces where participants "contribute, combine, revise, connect, evaluate, and integrate data and concepts" (Michelucci & Dickinson, 2016).

Efforts are underway to build large-scale online discussions about these complex problems, with real subject matter and the potential for practical implementation if the deliberation can produce good solutions. For example, Malone, Laubacher, & Fisher recently (2014) published a PBS Nova Next article titled "How Millions of People Can Help Solve Climate Change" highlighting the potential for crowdsourcing to solve complex challenges that no single person can possess enough expertise to fully address. A couple months later, the US government announced its intention to turn over full control of certain aspects of Internet governance, including its root name zone, to a global-scale multistakeholder process, if a proposal could be developed that acceptably provides for key governance functions (National Telecommunications & Information Administration, 2014; Pritzker, Crocker, & Chehade, 2014). After a years-long discussion and coordination process involving relatively frequent in-person meetings with thousands of attendees from around the world, the transition occurred on October 1, 2016 (Crocker, 2016). Even when that process tries to be very open to and welcoming of input from some of the previously uninvolved billions of people affected by the decisions coming out of this process, significant barriers to participation remain (Donahoe, 2016; Lange, 2014), cutting out a potentially valuable "long tail" of participation from a lot of people who might otherwise each be able and willing to contribute a little (Dunn et al., 2014). There is clear demand for processes and technologies that help enable large-scale online discussion around complex issues, which remains unmet.

[1] For a definitional discussion of what is meant by "wicked" problems, see also Rittel & Webber (1973).

Deliberative processes are often seen today in domains where many different stakeholders or stakeholder groups must come together and contribute their perspectives to arrive at a common decision. US Deputy CTO Nicole Wong, visiting CMU, described multistakeholder models in government as where "everyone brings their Lego block, and if we don't figure out a way and be incented to fit the Legos together, we build nothing. We all stick with our single block at the table…. Multistakeholder models, difficult as they are, we will catch more things with more brains at the table, that come from diverse viewpoints, than if we try to solve it solely as a technology issue, solely as a policy issue, solely as a legal issue, so having more people who are prepared to be open to other viewpoints and collaborating in that way is useful…. It is a skill set sorely needed" (N. Wong, 2014). Technology to support such forums, which can facilitate fitting the pieces together and supporting that kind of collaboration, could be valuable.

Technology cannot by itself produce the insights or results of deliberation, but act as a supporting platform to help aggregate, structure, and link together individual contributions so that human readers can form and benefit from emergent knowledge. In the words of Reddit founder Alexis Ohanian, "It's not the technology that does it. People are always the ones who are making it work. But now you're giving them a platform" (Schulman, 2013).

It has been shown that, under the right conditions, large diverse groups of people can arrive at solutions of better quality than any individual in that crowd (e.g. Howe, 2008; Page, 2008; Surowiecki, 2005). The canonical example of this is that a crowd's average guess for the weight of an ox at a fair (Galton, 1907) or a count of jelly beans in a jar is generally thought to be more accurate than even that of the best member of that crowd, when the task of accurate measurement is beyond the capacity of what any individual in the guessing population can or is allowed to do under the rules of the contest (Surowiecki, 2005). Prediction markets have been shown to be another reasonably effective way of aggregating individuals' opinions to produce more accurate predictions of the future than any individual

participant can reliably produce (Sunstein, 2006; Wolfers & Zitzewitz, 2004), though they still demonstrate a number of documented issues such as overpricing low-probability events (Wolfers & Zitzewitz, 2008), especially in thinner markets (Wolfers & Zitzewitz, 2004). These experiences demonstrate the need for input from a diversity of independent perspectives, as generally necessary conditions for the "wisdom of crowds" to operate (see (Page, 2008) for discussion of these and additional conditions). Diversity and expertise are both needed, with a proper balance depending on the application (Bonabeau, 2009; Phillips, 2014).

Online, it is possible to collect large and diverse collections of individuals' views. Ideation tools, as discussed below, are designed primarily to gather ideas as input from potentially large populations, allowing individuals to contribute completely independently of any content already in the system (e. g. when they initially come in), as well as to read, comment, and vote on others' ideas and submit new ideas based on new inspirations found either in others' ideas or outside the system. Generally, new ideas that are built on already-popular ones do not receive the same attention or number of votes. This may be because the already-popular idea that it improves upon may be the target of a greater number of circulating links, or because sites commonly sort proposals by popularity. When the base idea is already rising to the top of the list, especially if it lacks links to the improvement, even if a new voter could find both ideas, votes and feedback may be kept with the original so as to avoid "splitting the vote" and help that one idea get enough critical mass (which generally must be attached to a single submitted idea) so that it rises to the top of the evaluation list. Quality and ranking may not necessarily be correlated, under common system designs (Klein, 2009; Salganik, Dodds, & Watts, 2006).

Many social media technologies create more heat than light when applied to controversial, complex problems because they produce content that is often redundant, disorganized, polarized, and shallow, partially as a result of lack of support for navigating through content (Klein, 2011a). If the applied energy could be more coherently organized, these technologies might be better able to address the

targeted problem and make a difference. Efforts are underway to make listening and reflection more of a first-class activity in Web discussions, and tools to support this are desired (Kriplean, Toomim, Morgan, Borning, & Ko, 2012).

0.2.2: Collective Intelligence

At a minimum, we might hope that deliberation would allow groups to perform as well as the best member, as others are led to recognize the value of the best alternative, or perhaps to do even better than any individual in the group, by deriving new knowledge (Page, 2008), sometimes called the *assembly bonus effect* (Mercier & Sperber, 2011). When members of a community are providing the kind of listening, support, and scaffolding for one another described below, individuals who enter an online deliberation each with relatively little understanding emerge with a much greater understanding as a result of the group's collective activities, and the group as a whole may be able to demonstrate an "intelligence" beyond the maximum of any individual in it. This is sometimes referred to as "collective intelligence," which can increase the capacity of a group in a manner akin to a "general intelligence" factor (Woolley, Chabris, Pentland, Hashmi, & Malone, 2010).

Part of the value of supporting effective online deliberation is that it helps a group of people make collectively better decisions (Cavalier, 2011) because the group creates new knowledge in the process of effective deliberation (Mathews, 2014). Thus, design factors supporting effective online deliberation can play a part in supporting the development of collective intelligence, and design considerations relevant to collective intelligence can be worth considering in online deliberation systems, as one subset of tools that try to promote collective intelligence.

Successful deliberation with a large crowd of people involves synthesis and emergent wisdom, not just selection of a small set of "best" ideas. Even within an individual, human intelligence can be seen as an emergent property of billions of independent neurons linked together and interacting, which is similarly possible in large groups of interacting organisms, like an ant colony or a human society. "In a

spontaneous order, intelligence is not so much the property of each single component but rather something that appears 'between' all parts, making the whole bigger than the sum" (Liljeqvist, 2013).

In 2010, Malone, Laubacher, & Dellarocas published a "Collective Intelligence Genome" in the MIT Sloan Management Review, identifying a partial set of building blocks to give a deeper understanding of how collective intelligence systems work, to get at "the science from which the magic comes." The article presented building blocks with different answers to "What" (Create or Decide), "How" (Collection/ Contest/Collaboration for Create; Voting/Averaging/Consensus/Prediction Market for group Decide; Market/Social Network for individual Decide), "Who" (Hierarchy or Crowd), and "Why" (Money/Love/ Glory), along with partial information about "the conditions under which these genes are useful and the constraints governing how they can be combined."

The two "'How' genes" for Create tasks were described as Collection and Collaboration. Collection is useful when the activity can be broken into sufficiently small independent pieces; Collaboration is needed when manageable interdependencies exist (Malone, Laubacher, & Dellarocas, 2010). The Climate CoLab is an example of a project that uses the Contest subtype of Collection, though in reality the challenges presented by climate change have enough unavoidable interdependencies that the Collaboration "gene" may be needed, as illustrated by new "integrated contests" added a few years later (Malone et al., 2017). Much of the work in this thesis is aimed at the steps that can bridge from Collection [of contributions] to Collaboration, helping enable Collaboration while maintaining the ease of contribution and other benefits of the Collection gene.

Later conferences on collective intelligence followed up on different combinations of these "genes" ("Collective Intelligence 2012," 2012; MIT Sloan Newsroom, 2014). A literature review in the 2012 conference pointed out the need for a multidisciplinary investigation improving understanding about how micro-level attributes such as adaptivity, interaction, and local rules leads to collectively intelligent

macro-level emergent system behavior (Salminen, 2012). With better scientific understanding, we might be better able to design systems that "help society take advantage of the opportunities for organizing itself in new and better ways made possible by technology" (Hopkins & Malone, 2009).

In collective intelligence applications, framing effects are known to cause undue influences based on how a solution is presented (Bonabeau, 2009). In trying to implement "an application that taps into collective intelligence for improved decision making, …the devil is definitely in the details" (Bonabeau, 2009, p. 4).

Understanding those details, including the effects of structuring and framing content, is important to enabling implementation of systems that can effectively tap collective intelligence. Studies below, especially 3 and 4, explore this further.

0.3: Background: Explicit Formal Structuring and Linking Of Text in Online Deliberation Settings

In many platforms intended to support online discussions or deliberations around complex or potentially contentious issues in order to enable collective intelligence, the collaboratively produced artifact is a structured representation of the discussion itself. In these structures, there are no separate concepts of the content compared to the discussion behind it. This section reviews three such classes of platforms, still used today, which have played important roles in shaping perceptions about what such discussions might look like online.

0.3.1: IBIS and Argument Mapping

Some modern online deliberation tools have roots in the Issue-Based Information Systems (IBIS) pioneered by Kunz and Rittel (Kunz & Rittel, 1970). IBIS structures knowledge into topics, issues, questions of fact, positions, arguments, and model problems, with a designated set of possible relationships among these. Conklin and Begeman (Conklin & Begeman, 1988) developed a computerized graphical IBIS (gIBIS) system, for capturing small teams' design rationale. This hypertext

tool visualized IBIS's structured knowledge types (Issues, Positions, Arguments, and Other) as nodes in a network, with colors, filters, and other graphic cues to indicate node and link types and help designers use the IBIS model. This work evolved into the present-day Compendium (De Liddo & Buckingham Shum, 2010), Cohere (Buckingham Shum, 2008), and Evidence Hub (De Liddo & Buckingham Shum, 2013) tools.

Debate Hub continues the line of IBIS tools noted above, also combining elements of ideation and two-sided deliberation tools (see discussions below) (Shum, Liddo, Bachler, & Cornish, 2015). A variety of other tools have also elaborated and extended the fundamental concepts of IBIS. As one example, SIBYL helped adapt the model into more specific decision-making environments, e.g. by adding explicit representations of goals (J. Lee, 1990) in a decision matrix structure also used in later non-IBIS-based distributed-effort decision support tools (Kittur, Smus, Khamkar, & Kraut, 2011).

MIT's Deliberatorium (Klein, 2011b, 2011a) uses an IBIS-based structure to organize content, and adds features for rating, reputation, and user communication. Its **argument mapping** approach is intended to enable online deliberation that may have many different sides or cross-connections. The Deliberatorium has been used in topic-specific groups up to a few hundred people in size (Iandoli, Klein, & Zollo, 2009), though the use of argument mapping requires well-trained individuals.

Another particular kind of argument mapping is well suited for certain applications such as educational analysis of primary philosophical texts, leading to significant improvements in critical thinking skills (Harrell, 2005a, 2005b, 2011). Early reviews of argument mapping systems in these educational settings described requirements for making such systems effective, including being able to enter ideas (expressed in text) and move them around while retaining connections between them (Harrell, 2005a), in any order, consistent with a mapper's flow of thoughts (Easterday, Kanarek, & Harrell, 2009). This flexibility helps divide the work: one person may enter information and another may

organize it. Allowing users to enter text propositions prior to determining connections between them, thus separating the two steps and permitting different cognitive processes to guide each one, is valuable, but drawing all those connections is still a cumbersome manual task (Harrell, 2005a). A design lesson from Cohere (Buckingham Shum, 2008) teaches the importance of using an emergent rather than predefined structure.

Strictly-typed argument mapping technologies have taught us that an overly rigid structure can raise barriers to contributions. Especially in the early phases of formulating a set of contributions, a user may be unsure of whether each particular piece of knowledge is an Issue, Idea, Supporting Point, Opposing Point, or other specific allowed type. Forcing a user to consider the "type" of his contribution may prompt him to think more deeply about the content and how it fits in, but it interrupts the user's natural flow of thoughts and may be a frustrating requirement. Some users see the distinctions between different "types" of knowledge as arbitrary and ambiguous, with multiple plausible structures and the "correct" one not necessarily clear. Even the authors of these tools recognize that the primary costs of using them are associated with unbundling the contribution into its constituent elements and then locating the proper place to place these elements in the map (Klein, 2011a, p. 11); the costs of coming up with the contribution and typing it in are often negligible by comparison. If it is cognitively hard to classify and connect a contribution in a particular modeling scheme, there is a strong risk that it will not be recorded at all (De Liddo & Buckingham Shum, 2010, p. 6).

The current state of the art in managing this for argument mapping is to have one or several moderators verifying the content phrasing, type, and positioning before it becomes visible to non-moderators in the map (Klein, 2011a). A few individuals who have received special training in argument mapping are needed to enter or moderate the content, which creates a bottleneck for content, hinders realtime collaboration, and does not scale well, especially to maps that are larger and more complex than any moderator or moderator committee can manage. This is recognized as a key challenge and

explorations to overcome it include combining search tools with micro-tasks done by crowd members (Klein, 2011a).

Debategraph (Baldwin, 2010) improves on argument mapping with hyperlink exploration that does not require a single strict hierarchy, except in certain visualizations (as it offers multiple possible interface perspectives on the same debate content). Like the Visual Thesaurus (visualthesaurus.com), users visually navigate through a web of titled nodes, until they find the topic they are looking for, and drill down for additional information. However, this still requires an idea contributor to properly categorize the idea into one of the handful of options in a context-customizable ontology of IBIS-informed knowledge types (Baldwin, 2010), and connect it to at least one idea already in the map, as a barrier to contribution.

One of the key considerations for the design of online deliberation systems (Towne & Herbsleb, 2012) is to "Maintain low entry barriers for contributions of value." Low entry barriers are important for attracting new participants, especially from the "long tail" of many people each willing to contribute only a small amount, and increasing the efficiency of regular contributors' work. There is a large "cognitive surplus" of resources which may be available to help solve such problems, if we can figure out how to use them well (Shirky, 2010). People are looking to more actively engage with the world around them through e.g. sharing and creating content rather than just consuming it (Shirky, 2008); lowering entry barriers will be key to tapping into the "long tail" of diverse ideas which can have very significant impact (Klein, 2011a). Unfortunately, the manual structuring required by IBIS-based formats presents a somewhat high barrier to meaningful participation which, while lower than what had existed previously, is still too high to be suitable for widespread adoption. Steps that might help with this include separating the content generation from the steps of content categorization and organization or providing a view focusing on conclusions if a discussion has reached any, which may provide an easier point of entry than trying to digest a large nonlinear map. The separation of conclusions from

discussion, however, may lead to changes in readers' expectations and understandings of the system, community, and/or content produced. These changes are experimentally explored in much more detail in studies 3 and 4 below.

0.3.2: Two-sided Deliberation

One of the key weaknesses in earlier work on analysis and support of deliberation and debate is the over-simplified concept of issues as having two discrete sides, characterizing positions as either pro or con. Some summarize the discussion with a visual indicator of the balance between the two sides. This is the approach taken in deliberation interfaces like Debatewise, Debatepedia (later Debatabase), ProCon.org, and others (Lindsay, 2009), including those that were still being built concurrent with this thesis (e.g. Miniter, Kriplean, & Toomim, 2014; Pitsos, 2016; N. Santillo, 2013; T. J. Santillo & Santillo, 2013). This approach is also visible in language technologies for analysis of stance that take as their goal to represent how texts reflect one side or another (e.g., Malouf & Mullen, 2007; M. Paul & Girju, 2009; Yu, Kaufmann, & Diermeier, 2008), or those that focus just on positive/negative sentiment analysis of e.g. comments on position articles. However, complex issues often do not have just two exactly opposed sides. Instead of black-and-white issues, there is not only a lot of gray but also many shades of every color and color mix, with unclear boundaries between concepts. Effective decision making on complex issues requires insight into the direct and indirect ways alternative choices affect a diverse multitude of stakeholders. The necessary weighing, balancing, and satisficing that is inherently part of deliberation (Mathews, 2014) can only be accomplished by drawing on rich problem representations that go beyond simple, two-sided representations of issues, and pro versus con representations of arguments.

With a proper treatment of the multifaceted nature of issues for deliberation, the contributions may be better organized to highlight areas that a given reader cares most about, which is motivating for the readers and participants (Kraut & Resnick, 2012). This makes the collection more usable and navigable

while still respecting the "fuzziness" and uncertainty in conceptual boundaries, which do not clearly exist. Two-sided deliberation is easier to navigate, but does not allow the field of online deliberation to be appropriately applied to complex problems where the large diversity of perspectives it could harness are most needed.

0.3.3: Ideation Tools

Ideation and idea management systems, where many individuals submit ideas in response to a particular prompt and can (usually) vote other ideas up or down (and sometimes comment) are popular and growing. Competitive providers each report millions of users and hundreds of thousands of submitted ideas (IdeaScale, 2013). Microsoft, IBM, Dell, Whirlpool, and UBS have used this approach to tap the knowledge of their employees and customers (Bailey & Horvitz, 2010). The International Conference on Communities and Technologies also specifically focused a workshop on large-scale idea management and deliberation systems (Convertino, 2013).

One common process for using these tools is to have a diverse crowd first generate many contributions through some open brainstorm-like process dominated by generation and divergent thinking, followed by a review and sometimes narrowing of those ideas. The convergent data reduction stage is made easier by tools for organizing and grouping or connecting related ideas, usually based on the ability to measure similarity. Sometimes, there is no specific timeframe distinguishing divergent and convergent phases, which may happen in different parts of the discussion space simultaneously. Capacities for such simultaneity may be especially important in disaster response or other fast-moving, time-sensitive applications, as discussed below (p. 32).

On many platforms, the convergent data reduction stage is done simply by aggregating vote counts. Even where the best or most representative ideas are separated from others by expert review, factors which influence the way readers perceive quality in the contributions influence the outcomes of ideation processes on these platforms.

Processing through such ideas, the aim is often data reduction, such as finding the best ideas, or identifying cross-cutting concerns or themes that may be present in many ideas. Another key aim is discovering deep knowledge embedded in the data, and not just the answers to questions a user thought to ask (Nahm, 2004, pp. 1–2). Synergy between already-contributed ideas, if discoverable, can lead to new and better contributions, and is arguably the source of most good ideas (Johnson, 2010). Contributions of Study 2 below allow us to better evaluate and understand the limits of automated support for discovering those potential synergies.

In the absence of a good approach to navigability, we see that enormous and often impractical effort is required to "harvest" knowledge from large sets of contributed ideas. For example, an organizational health forum at Intel elicited 1000 posts from 300 participants, and the post-discussion analysis team spent 160 hours (≈10 minutes/post, perhaps longer than they took to write) going through the content. They found "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" (Klein, 2011a, p. 2). Google's "10 to the 100" project fielded 150,000 ideas from over 170 countries, and the 3,000 employees recruited to go through them were still nine months behind schedule (Klein, 2011a).

At a town hall meeting announcing major changes to Quirky, a site for crowdsourcing consumer product ideas, the CEO was asked "Why are you not searching the archives for similars and things like this?" He responded by saying that "It is absolutely impossible for humans, real human beings, to read 300,000 ideas, scaling by several thousand per week, and try to interpret different people's use of the English language, … and try and understand 'is this is something that relates to something we've seen?'" Although he would love the company to be able to do that, they just can't (Kaufman, Ben, 2015). This is a source of significant frustration for some idea proposers (Njaa, 2015).

Based on his experience with a 12,000-employee company managing text data in a specific domain, the chief scientist of M*Modal describes the key challenges when working with large volumes of data as aggregating the pieces, relating them to one another, boiling them down, and intelligently sifting through the data to identify the needle in the haystack relevant to the current situation (Fritsch, Pesenti, Hutchison, Cancelliere, & Srini, 2013). Bailey & Horvitz (2010) specifically note that comments on ideas commonly try to link the author/idea and other people, efforts, or ideas related to the one posted, that these connections are an important but often unmet expectation motivating contribution, and that automatically attaching this kind of link information to ideas could be useful. Study 1 below explores these questions and motivations in much more detail.

As another example, IBM hosts online "Jams," online events over the course of a few days to solve business challenges, clarify company values, and produce ideas for new initiatives or improved operations. A 2006 Innovation Jam, described as the largest-ever online effort to advance technological innovation, involved 150,000 people from 104 countries, investing $100 million in the best ideas from the event (Bjelland & Wood, 2008; IBM, 2012). It generated billions of dollars in revenue, followed by another Innovation Jam in 2008 and spin-off consulting service (Bjelland & Wood, 2008; Cleaver, 2013; IBM, 2012). Primary goals for that Jam included connecting people in an exciting way that helped them build on each other's ideas and create something new and innovative (IBM, 2012). The actual experience, however, fell short: contributors were not constructively building on each other's postings, and new visions and connections did not emerge until the manual review process after the event (Bjelland & Wood, 2008).

This review process required teams of specialists and senior executives to spend weeks sorting through gigabytes of text, to pick a few new ideas from tens of thousands of postings. This mostly-manual process did produce business success, helping IBM listen to already-circulating ideas and helping executives combine related ideas in major new initiatives. However, extracting value from the ideas

required a great deal of management time and participants were not readily able to find and build on one another's ideas, or connect directly to implement them (Bjelland & Wood, 2008, p. 37).

After Japan's triple disaster (earthquake + tsunami + nuclear meltdown) in spring 2011, IBM adapted their Jam to an issue-specific forum which drew voluntary contributions from 275 employees in 23 countries (20% of the 1250 employees from 45 countries who actively registered) on seven topics, each at the root of a tree-structured text-based discussion, with an average of 100 responses per topic, over the course of just a few days (Muller & Chua, 2012). Leveraging the power of distributed knowledge and efforts to respond to natural disasters is becoming increasingly common (Goggins, Mascaro, & Mascaro, 2012) and necessary (IPCC, 2014), and requires computing innovation with empirical grounding toward rapidly, dynamically structuring information contributed after such events, so that contributors and officials can more easily make sense of the situation and act appropriately (Palen et al., 2010, p. 6). Disaster situations generally do not afford months of manual review, and no specific manager may be clearly responsible or capable of doing so. Navigability is a key attribute in successful disaster relief discussion sites (Goggins et al., 2012, p. 58; M. J. Paul, 2001). Generating navigation links dynamically, as explored in Study 2 and to some extent Study 1 below, helps rapidly incorporate new information, which may arrive in the form of new documents and confirmation or disconfirmation of pieces of information or relationships between them.

Social media and ideation tools can generate a lot of activity, but only a small percentage actually goes to solving the intended challenge (Klein, 2011a). After a certain point, the amount of truly new content and perspectives grows more slowly than the number of participants (Klein, 2011a), especially if the new participants do not make the crowd more diverse (Page, 2008; Surowiecki, 2005).

The potential impact of these systems is great, and ideation systems can help gather many contributions together, but certain constraints such as navigability limit their present usefulness. With

all these contributions, how do you navigate around and read through tens of thousands of ideas efficiently? "Extracting and using the good ideas of tens of thousands of people is not simple, but it is potentially powerful…Idea generation is in some ways the 'easy' part of innovation, whereas advancing, refining, and building support for those ideas is the really tough part," according to IBM's VP for technology and innovation programs (Bjelland & Wood, 2008, p. 40). Further, to the extent that participant perceptions of idea quality determine outcomes, what factors affect those perceptions and how might these factors be influenced by design decisions, e.g. around how closely to connect an idea with comments or discussion posted to it? The studies in this thesis help address those questions.

0.3.4: The Problem of Navigation

Deliberation may be able to happen online, where individuals in a large and diverse population can access a platform and contribute even a small amount, if provided the right affordances and supports for doing so. A number of existing online deliberation tools and their shortcomings are discussed above. After a review of many online deliberation systems, one set of the core considerations for systems moving forward focused on design for navigability (Towne & Herbsleb, 2012). That is, once an online deliberation site has attracted contributions, they should be organized so that people seeking particular information can quickly find it, and potential contributors can easily locate where their contributions fit in. Design considerations include the following:

Organize content topically, rather than temporally. This makes it easier to locate specific topics of interest. Current online deliberation approaches that use e-mail lists, Web forums, blogs, or comment chains on blog posts and news articles often organize content according to the sequence in which it was added. This can make specific contributions hard to re-locate, and prompts some contributors to repeat their points many times. It also makes finding topical connections to trace conversation threads difficult, although some attempts have been made to automate the process (e.g. Y. Wang, Joshi, Cohen, & Rosé, 2008).

Topical organization works well in allowing people to find places to contribute to Wikipedia, for example, and is contrasted directly with the challenge of finding discussions within Talk pages on the same platform, especially the longer ones. For example, the talk page on Barack Obama was more than double the length of *War and Peace* even before he became President, and nobody was reading through all that content to find previous discussions before restarting them even when they were shorter, much to the frustration of participants in the prior discussions (Vozick, 2006). Individuals interact with online discussion very differently when it is organized topically rather than temporally, finding it easier to jump in to various points of the conversation and address multiple aspects of an issue that can easily get lost in a temporally-based thread structure (Y.-C. Wang, Joshi, & Rosé, 2008).

Minimize or eliminate duplication. An understandable topical organization structure is necessary but not sufficient to minimize duplication. Ideation tools, one-way interaction tools such as Regulations.gov, and even discussions on some structured platforms like Cohere (Buckingham Shum, 2008) have issues with significant duplication of content, often because similar content cannot easily be discovered, linked, or merged together. Some community support systems (e.g. Facebook Help), general question-answering sites (e.g. Stack Exchange), bug trackers (e.g. Bugzilla), and ideation tools (e.g. IdeaScale) incorporate search utilities to help prevent duplication at the time of posting, which help but do not fully solve the problem, nor do these utilities fully address the challenge for readers. These features don't easily allow collaborative refinement, and are generally unable to identify cases where different terms refer to similar concepts. The position of each contribution within the content structure should also be modifiable, so that users can adjust the structure as it evolves.

Relate solutions to one another. Designs should facilitate the creation of links between reasons and elements of an argument that make logical and semantic sense, and allow the user to view multiple topics simultaneously, so they can compare and bridge knowledge across them (Easterday et al., 2009). These links should not be restricted to a hierarchy or other rigid structure. This point is missed by online

debate technologies that enforce a strict pro/con format, as described in section 0.3.2 above. Cohere (Buckingham Shum, 2008) demonstrates that links can and should be more expressive than simple URLs. Adaptive automated tools can help extract tacit structure and uncover patterns of relationships between topics and between participants. This could help reveal solutions as well as "holes" where there are prime opportunities for particular individuals to contribute.

Provide rich opportunities for exploration. Once concepts are related to one another in the online deliberation system, visitors should be able to explore the network via hypertext links and visual exploration. Examples of this principle can be found in gIBIS/Compendium (Conklin & Begeman, 1988; De Liddo & Buckingham Shum, 2010) and DebateGraph (Baldwin, 2010). In the DebateGraph default view, for example, users visually navigate through a web of titled nodes, until they find the topic they are looking for. They can hover over a node for a short description, and click on it for more details. Hyperlink exploration should be one of several options for navigating through content. This is done well in Wikipedia, via "wikilinks" reaching articles about terms and concepts linked to from where they are referenced elsewhere.

Unfortunately, readers navigating platforms like Debategraph also have an arbitrary limit of how much state and history (e.g. of what they have seen) they can maintain at once. An alternative design, which better supports working sets comprising various user-chosen snippets, is demonstrated in Bubbles (Bragdon et al., 2010). The Bubbles interface design was motivated by observing the challenges of navigability within a code base, including measurement of developers spending on average 35% of their time navigating within and between source files and an equal amount of time on failed searches, in software maintenance and enhancement tasks on an unfamiliar but small code base (Ko, Myers, Coblenz, & Aung, 2006). Although the task is not necessarily the same, potential contributors exploring an online deliberation space may face challenges similar to those discussed in (Ko et al., 2006) in keeping track of and returning to relevant snippets they've found. Problems such as vocabulary diversity or

inaccurately descriptive titles seem even more likely to be issues in online deliberation than in software development (as pointed out by participants in Study 1). The same interface concepts demonstrated in (Bragdon et al., 2010) may be useful in exploring online deliberation spaces, for many of the same reasons.

It's been said that "if the content isn't structured logically with a simple flow, it might as well not exist" (Bank & Cao, 2015, p. 36; Cao, 2015). At the least, effective content organization and navigation by definition makes it easier for readers to find content of interest. This increases the potential audience size and the value perceived by contributors who are motivated to contribute something valued by the community, which may encourage further quality contributions (Klein, 2011a, p. 11).

0.4: Background: Navigation through Textual Corpora

The previous section discussed some limits of current systems and barriers to participation that can be presented by requiring potential contributors to carefully place their contributions, focusing on a few types of "traditional" platforms for online deliberation that cover a range of degrees to which content creation is separated from the step of placing a contribution in the network of ideas around it.

Platforms for collaborative work involving textual contributions from a large diversity of contributors are becoming increasingly prevalent and productive, whether it be in areas like online deliberation, discussions around collaboratively edited products such as Wikipedia articles or open-source software patches (Tsay, Dabbish, & Herbsleb, 2014), blogs as a venue for large-scale public discussion (Sunstein, 2006), forums, venues for social support (Dinakar et al., 2012; Hwang et al., 2010), or ideation tools. An increasingly large volume of information, work, and communication is based on, accomplished through, and/or represented by collections of text documents. These collections are increasingly often too large to read through them one document at a time. It can be extremely difficult for users, especially new users, to find desired information on such platforms, especially when the users themselves do not understand enough about the corpus or the pieces they would find most interesting to specify a

keyword query in a search engine. This creates a pressing need for tools supporting navigation within and analysis of text collections (Candan, Di Caro, & Sapino, 2012, pp. 1–2; Cui, Qu, Zhou, Zhang, & Skiena, 2012, p. 1; Gretarsson et al., 2012, p. 2; S. Liu et al., 2012, pp. 1–2; Zhao et al., 2011, p. 2). This section examines a relatively broad history of approaches to enhancing navigability through large document corpora.

0.4.1: Link-based Navigation through Textual Corpora

After World War II, Director of the Office of Scientific Research and Development Vannevar Bush challenged scientists to increase accessibility to the inherited, distributed store of knowledge humanity has accumulated, as their top scientific priority after emerging from war. By that time, capacities for generating new information exceeded capacities for accessing existing information (Bush, 1945). Bush's prescient essay arguably (Conklin, 1987; Shalizi, n.d.) resulted in the later invention of hypertext and accurately anticipated many other significant advances in technology. In it, he highlighted the "artificiality of systems of indexing" and the difficulty of having to navigate through volumes of records in hierarchical classification trees, clearly distinguishing this from the association-based web of concepts in human thought. The 1945 essay describes a vision for automating "selection by association, rather than indexing" similar to how a mind follows an associative trail, tying pairs of items together so that "any item may be caused at will to select immediately and automatically another" as the essential feature of then-future technology (Bush, 1945). "Technical difficulties of all sorts" left implementation of this vision to the future. In the following decades, large parts of the vision were realized through electronic storage of texts, increasing data densities on storage devices, and transmission over early computer networks, but the most essential and defining component of this vision proved to be very challenging.

At CERN in the late 1980s, as at other times and with other organizations, a large volume of text data and documentation was being generated and stored electronically, but was still not very well

mentally indexed especially in an organization with a lot of turnover. When needs arose for a local change to e.g. some technology or protocol, it was difficult to find what other parts and people would be affected. The contents and structural interdependencies between components made maintaining a single information source impractical, and Tim Berners-Lee saw that these kinds of problems would soon become more widespread (Berners-Lee, 1989), as they did with the Web as "the prototypical example of a big knowledge organization problem," attempting to organize and represent large quantities of information for an unknowably large and diverse audience (Mai, 2010). He proposed "a 'web' of notes with links (like references) between them" (Berners-Lee, 1989). This resonated with the growing community familiar with the concepts of hypertext, as surveyed a few years earlier (Conklin, 1987).

That proposal, which became the World Wide Web, was aligned with a publication from UCLA in the same year, describing users' navigation through online information spaces with an analogy to "berrypicking," proposing a shift from prior work that focused on matching queries with controlled-vocabulary document representations, to an iterative process where users iteratively refine their queries including changes in what they are seeking, based on learnings from early query results (Bates, 1989). This suggests that it is important to not only allow evolving queries but also design to support users' efforts to browse and to navigate between related pieces of information within a database. (Browsing and berrypicking are separate concepts in that work). That work specifically discusses content relationships such as forward and backward citation chaining, use of manually-created topically-based indexing services, querying the physically adjacent areas of a library (assuming content-based clustering), or finding other works in the same venue or by the same author (assuming some topical consistency in those sets) (Bates, 1989). This paper recommended several specific "key design features" to better expose these relationships (see esp. pp. 11-14), many of which are now common in repositories like the ACM Digital Library, and specifically called out the potential for hypertext to support the berrypicking approach to information retrieval (Bates, 1989). Link structure has also been

used as a basis for finding related Web pages "where the input to the search process is not a set of query terms, but instead [the URL of] a page" (Dean & Henzinger, 1999).

0.4.2: Keyword-Based Navigation through Textual Corpora

Early hypertext systems distinguished between hierarchical and referential links, stating a clear need for cross-hierarchical links and a shortcoming that "the creator of a hierarchical organization must anticipate the most important criteria for later access to the information" (Conklin, 1987), which is difficult. Keywords were also considered to provide an important source of linking (Conklin, 1987). Berners-Lee discouraged the use of hierarchical trees for navigation and pointed out that keyword-based information access is problematic because keyword choice is inconsistent, especially between people (Berners-Lee, 1989). In the real world, items may also not fall naturally into cleanly separable categories (Mai, 2010), and sometimes the most interesting, promising, and fruitful parts of a discussion are ideas at the intersection of multiple topics and sources (Johnson, 2010). This can be helped by using a non-hierarchical system that permits multiple tags, such as the keyword-based navigation leading use in some large textual corpora such as StackExchange question-answering sites. Tagging or categorization is sometimes done by the contributor at the time of the submission to idea management systems (e.g. Bailey & Horvitz, 2010), but this suffers from *polysemy* (a tag may have many meanings) and *synonymy* (different tags may mean the same thing), and requires that the contributor know which tags may be appropriate in the corpus and take the time to apply them. Together, these problems are referred to as a *vocabulary gap* and are endemic to traditional keyword search (Boytsov, Novak, Malkov, & Nyberg, 2016). In a study of the problem across a variety of domains, the probability of two people using the same word to describe a given concept was found to be less than 20%, leading to an 80-90% rate of failure for keyword-based access (based on *synonymy*), while the probability of two people referring to the same concept with a given word ranged from 13-73% depending on domain (the

polysemy problem) (Furnas, Landauer, Gomez, & Dumais, 1987). In approaches that suffer from these problems, it is difficult to find, connect, consider, and/or synthesize ideas based on topical relationships.

Keyword-based knowledge organization systems "cannot be objective but instead take on the bias and perspective of their designers" (Bullard, 2014). Classifications are all arbitrary and conjectural to a certain degree (Mai, 2010); even the *ideal* of a "perfect" ontology is arguably a mistake (Shirky, 2005). Berners-Lee proposed that "documents on similar topics are indirectly linked, through their key concepts" (Berners-Lee, 1989, p. 8). He also states that "Much of the academic research is into the human interface side of browsing through a complex information space. Problems addressed are those of making navigation easy, and avoiding a feeling of being 'lost in hyperspace,'" but more fully addressing these questions is deferred until greater technical maturity (Berners-Lee, 1989, p. 14). Having reached a time of greater technical maturity, it is time to revisit these questions.

An early hypertext-like attempt to allow readers to navigate through a document using keywords relatively free from the constraints of the original author's organization, was SuperBook, which took as input a long structured nonfiction text (e.g. reference book) and added navigation support such as keyword search with stemming-based query expansion. Superbook helped users identify potential instances of polysemy by indicating how many search results were found in each section or subsection on a table of contents list, and helped users deal with synonymy (once discovered) by specifying synonym sets that would then be automatically used in query expansion for searches on any query in the set, including by other users. It is labeled hypertext"-like" because it lacks support for author-generated machine-followable links, but not requiring that the author create all those links can be an advantage increasing the general applicability of the technology (Remde, Gomez, & Landauer, 1987).

This kind of system has advantages for readers trying to answer questions not anticipated by an author's organization of a text. However, it left open the problems associated with finding information

when the searcher is not using the same words used by the author, and fruitless searches or false leads in those cases presented disadvantages and poorer search performance compared to printed (paper) versions of the text (Egan, Remde, Landauer, Lochbaum, & Gomez, 1989). This led the same research group to invent Latent Semantic Indexing (LSI) / Latent Semantic Analysis (LSA), as an intended solution to the problems of synonymy and polysemy inherent to keyword-based approaches to navigability (Dumais, Furnas, Landauer, Deerwester, & Harshman, 1988). LSA is a conceptual predecessor to the LDA approach discussed in greater detail below.

0.4.3: Cluster-Based Exploration of Textual Corpora

An extension of keyword-based navigation is cluster-based exploration, where keywords or summaries for each cluster are automatically generated. These clusters can then support navigation and the discovery of desired information on sites with large volumes of text content (Zhao et al., 2011). A landmark example (Carpineto, Osiński, Romano, & Weiss, 2009, p. 11) is the "Scatter/Gather" framework which helps a user explore a document corpus without knowing enough about the vocabulary or target document to formulate a search query (Cutting, Karger, Pedersen, & Tukey, 1992). As originally described, this method uses a clustering-based approach to search, as opposed to a nearest-neighbor search, assuming that documents each belong to a single topically-themed cluster (Cutting et al., 1992). At each focus-narrowing step of Scatter/Gather, a user selects a subset of clusters determined and summarized by some clustering and summarization algorithms, which re-run over the subset of documents in those user-selected clusters (Cutting et al., 1992). This method seems to be a less effective standalone information retrieval tool than keyword search (Pirolli, Schank, Hearst, & Diehl, 1996), but cluster-based exploration in combination with other methods such as search has been shown to be a valuable strategy (see e.g. Carpineto et al., 2009; Hearst & Pedersen, 1996; Koshman, Spink, & Jansen, 2006).

When Scatter/Gather was introduced, evaluation was deferred awaiting "metrics appropriate to the vaguely defined information access goals" it was designed to support (Cutting et al., 1992, p. 325). Though a few users were generally able to choose clusters that contained the highest number of documents TREC-rated as relevant to a given query (Hearst & Pedersen, 1996), it is "not easy to assess the knowledge about a collection that is communicated by a browsing technique" (Pirolli et al., 1996) and user-based evaluation of cluster-aided search was still "largely an open question" as of 2009 (Carpineto et al., 2009). Methodological contributions made by Study 2 below help address this gap.

0.4.4: Thread-Similarity-Based Navigation through Textual Corpora

Xu and Ma (2006) attempt to determine topical similarity of forum threads based on their content, in order to make the content more accessible for all Internet users. They cite the large percentage of automatically-created links between forum pages as an obstacle to using link-based algorithms like PageRank as an aid to finding content of interest, and remove those links from the network as a critical preprocessing step (Xu & Ma, 2006, sec. 4.2). Xu and Ma assume that all documents can be classified into a unique topic within a topic hierarchy on the basis of a few discriminative words, which differs from the assumptions made in this thesis that documents generally have a mix of topics and that topics are not "naturally" organized in a hierarchical tree (Xu & Ma, 2006). Xu and Ma (2006) apply a topic-specific "personalization vector" to bias the PageRank algorithm, assuming PageRank's random but task-oriented surfer is more likely to jump to a page that is topically similar to what he is reading currently (when not following links), without explaining how the surfer would find those pages.

Xu and Ma (2006) evaluated their work by choosing what appeared to them to be frequently asked questions from forum sites, and formulated search queries of a few words each from these. They used a BM2500 bag-of-words ranking similar to TF-IDF and slightly more tunable than the older BM25 ranking (see (He, He, Gao, & Huang, 2002, p. 467) for details) to select the top 100 pages from those queries. They then used their content-based ranker in addition to traditional PageRank to select the top 30 pages

by each method. The authors judged the pages ranked more highly by their content-based ranker to be more related to the queries, claiming a significant difference on one of two measures used. Study 2 below contributes a methodology and instrument that could be used for more robust evaluation of topic-based similarity measures.

Singh et al. (2012) subsequently attempted to address navigability between threads in online forums, "finding threads that are similar to a given thread" and adding links between them. One of their motivating applications is for "knowledge authors" to be able to cover one type of problem fully before switching contexts to the next. After discussing the importance of accurately measuring lexical similarities of discussion threads, their contribution focuses on ways of leveraging the structural information in the multi-author exchange, such as overweighting the title and first post. Their evaluation is based on about 45 hours of humans' labeling whether or not candidate threads were relevant to 38 query threads (about 85 candidates per query). The paper doesn't say how the queries or candidates were chosen (from the Apple Discussion Forums).

The primary similarity measure Singh et al. tried was TF*IDF cosine similarity (2012, sec. 5.2) and they also used an LDA model (2012, sec. 5.6.2) but they don't report a direct comparison. They also say that "Selecting the right number of topics is an important problem in topic modeling," then arbitrarily choose 100 topics (results aren't reported) and show that the performance isn't sensitive to number of topics in the range 1 to 7. Study 2 below presents one solution that can help choose this parameter (also used in Study 4), which in this thesis provided a relatively smooth function for model selection.

0.4.5: Human-Recommendation-Based Navigation through Textual Corpora

Another approach to navigation through large text corpora is to use the services of a human guide who has heavily invested in corpora-relevant expertise. For example, in libraries, experts train for years to become familiar with the content of their collections, to be able to recommend materials to inquiring readers. This curation service is still alive and well and being adapted to new interface technologies (e.g.

by using the Web to make and respond to requests). As the size and diversity of the challenge or requesting population grows past a certain point, this model breaks down because "librarians do not scale well" (Leber, 2014), though specialization can help handle subject matter diversity and adding staff can help handle request load. Yahoo!, which started "as a directory of websites that helped users explore the Internet" (Rossiter, 2014), closed that service 20 years later (Sullivan, 2014), perhaps recognizing the limits of scalability in a human-curated directory of the Web.

Automated book recommendations "work better for certain categories of books, such as topical nonfiction" (Leber, 2014, p. 5), suggesting this would be a better domain to start with in training a content-based connection discovery engine. Even though automated recommendation engine Amazon Goodreads requires users to rate 20 books before personalizing title recommendations, these recommendations do not seem to work very well in fiction (Leber, 2014). Users find greater accuracy and value in "simply knowing there was a human being involved…adding an extra level of curation…between machines and people, it all goes hand in hand." (Leber, 2014). This suggests that involving humans in the curation of conceptual links is seen as valuable.

Card sorting, a technique for hierarchically structuring Web sites (and organizing other information spaces, after being used in the Web (Hudson, 2014)) based on users' conceptual connections between predefined elements, helps uncover terminology, relationships, and categories to help organize largely static information (Hudson, 2014; U.S. Department of Health & Human Services, 2013), subject to some of the same limitations in term ambiguity as described in section 0.4.2 above. There are even service businesses based on the value in uncovering this information (e.g. Optimal Workshop, 2015; Webcredible, 2015).

While expert human guidance is useful, this approach is limited in scale, scope, and speed, especially with regards to incorporation of new contributions to the corpus. Even "discovery-oriented workers"

who have undergone long apprenticeships to become experts in their field do not necessarily develop recognition-based expertise (like that of a chess master) because discovery happens at the frontiers of knowledge where nobody is an expert (Valdés-Pérez, 1999); this is especially likely to be true in 'wicked' problem spaces where the sheer complexity of the problem means that nobody can fully master it (Rittel & Webber, 1973).

0.4.6: Algorithmic Link-Based Navigation through Textual Corpora

Another way to improve navigability, explored in greater depth in this thesis, is to organize the expression of issues and ideas into sets of topics and then help provide structural hyperlinks between content elements based on relatedness measures such as similarity. Even an approximate nearest-neighbor search based on a similarity function that can take into account subtle term associations has proven to be generically effective especially for textual data (Boytsov, 2016; Boytsov et al., 2016). Although the size of content elements varies from one setting to the next, an appropriate length for the unit of analysis, or node in a navigational web, is in many applications one to several paragraphs (Conklin, 1987).

The context of an interaction is known to affect how people formulate contributions to that interaction. There may not be a simple notion of what counts as "similar" for this purpose, an issue explored in Study 2 below. There is also limited prior understanding about how showing interaction as part of a context affects perceptions of the "current" or "focal" element, and this gap is addressed in Study 3 and Study 4 below.

Prior work in attempting to automatically generate similarity links between documents in a text corpus reduces the dimensionality of "bag of words" text vectors using, for example, variants on Principal Components Analysis (PCA) or Latent Semantic Analysis (LSA) to help structure large text corpora in order to improve navigability (LSA e.g. Dumais et al., 1988). Work in this thesis (esp. Study 2 and Study 4) builds on newer models of text, namely Latent Dirichlet Allocation (LDA), developed

explicitly to enable efficient processing of large text corpora for similarity and relevance judgments (Blei, Ng, & Jordan, 2003). Models like LDA are often used to reduce dimensionality to themes that are useful building blocks for representing a gist of what a collection contains, statically or over time (e.g. S. Liu et al., 2012). Without validation against human perceptions of similarity, the algorithm's usefulness for increasing navigability in large text collections by connecting documents that have similar mixtures of topics has been asserted (e.g. Steyvers & Griffiths, 2007). For example, LDA has been used to find connections between stories written by distressed teens, helping authors see that they are not alone in their plights (Dinakar et al., 2012). In a 12-person evaluation, LDA selections were seen as closer matches and more helpful to original authors than those based on TF*IDF (Dinakar et al., 2012).

This use of LDA makes sense, however, only if the representation of texts that it yields enables algorithmic similarity computations (such as cosine similarity) having results that closely resemble similarity as experienced by human users when they compare the original texts. Yet many of these technologies are based on measures of similarity between documents without clear validation that the algorithmic measures being used match human perceptions of similarity. Measuring how well various similarity measures match human perceptions is important when those measures are being used to help humans navigate through a large set of elements, e.g. in online communities (Spertus, Sahami, & Buyukkokten, 2005). Understanding the limitations of a model also helps inform applications of it.

Observation that statistical tasks such as unsupervised classification are "still largely based on ad-hoc distance measures with often no explicit statistical justification" has previously motivated work exploring statistical properties of those measures and mathematically-demonstrated suitability of use for certain applications (e.g. El-Yaniv, Fine, & Tishby, 1997; Lin, 1991). Analogously, tools supporting human tasks such as navigation in a text corpus are still largely based on computed distance measures with often no explicit experimental validation against human perceptions. Study 2 shares a method for such testing as well as results for a commonly-used measure (LDA/cosine similarity).

Because the set of features chosen to represent the text is the only basis for determining how similar or different two texts are, this set of features (representation of data) both enables and limits what kind of structures and connections can be found in the data. In all cases, some information is lost from the original text, and other aspects are emphasized or deemphasized. For example, topic vectors may enable connections to be found between pairs of documents that use different words when describing the same concept, while alternative representations such as vectors containing psychological dimensions of words can help find other similarities, such as those based on psychologically common authorship (Boyd & Pennebaker, 2015). The choices involved in expanding or collapsing (e.g. stemming) the text prior to processing impacts the speed and accuracy of similarity measures computed from them (Metzler, Dumais, & Meek, 2007), as does the metric chosen to operationalize similarity based on that representation (Mohler & Mihalcea, 2009). Approaches to measuring text similarity generally measure closeness of distributions of text features, size of angle between vector representations of text, or distance between end points of vectors within a vector space. Being able to analyze "errors" or ways in which the algorithmic relatedness measure fails to match human perceptions, as well as times when it does, will enable improvements on later iterations. Study 2 below contributes to such measurement.

Relational Topic Models (RTMs) for document networks (Chang & Blei, 2009) model documents and the links between them in a shared latent space of topics, with documents represented by standard LDA topic distributions, and links represented by distributions over the same LDA topics, numerically computed using the element-wise product of the documents' topic vectors. RTMs can be used to analyze linked corpora such as citation networks, hypertext, and explicitly networked user profiles (e.g. in social media). One of RTM's novel contributions is that it can predict new links given words in a document new to the network, without assuming a separate vocabulary for the links; RTMs constrain a document's words and its links to be explained by the same topics. The work in this thesis returns to the

LDA origins of this model (i.e. LDA without other explicit connections between documents) looking at the use of these concepts on data sets that do not have links between documents to begin with.

The need for technologies that can automatically detect conceptual connections between text documents is evident in the marketplace. After his success selling Blogger to Google, Twitter co-founder Evan Williams started a blogging platform called Medium, based at least in part on his perception of the problem of information overload and a need for gatekeepers, but "one that relies heavily on technology rather than human expertise or taste. While it has some editors soliciting and promoting some content, the bigger idea is to use algorithms to help identify blog posts that readers consider valuable and to bubble them to the surface… We want to create the potential for great things to be created and found." Work started in 2012 and opened to the public in October 2013. The recommendation algorithms are based primarily on exposure metrics like which stories are read and e-mailed the most, voted up at Reddit, or Liked on Facebook (Richtel, 2013). However, algorithms that make popular content even more popular (e.g. sorting by upvote count) can dissociate popularity and content quality (Salganik et al., 2006), and even Williams admits that "I can't claim I've figured it out." Having more effective support for navigation, which this thesis seeks to contribute to, would help fulfill this identified need.

During this thesis work, Wikipedia released an "Explore" feature for its iOS app recommending top rated and random articles as well as those "recommended based on what you've read" (Minor, 2016), to make it easier to better discover content matching users' interests (de Looper, 2016; Perez, 2016), incorporating the previously external content discovery functionality of a third-party app Curiosity (Underwood, 2015). Wikipedia's fairly dense network of manually created wikilinks can help in determining what content is related to what others, though it is not clear exactly what signals are being used in this app (Minor, 2016).

Hewlett-Packard's paper "Finding Similar Files in Large Document Repositories" (Forman, Eshghi, & Chiocchetti, 2005) cites customer satisfaction improvements resulting from identification of similar files (overlapping a lot in textual content, such as duplicates or different versions of the same document). Their approach applies a sliding window over each document in the set, computing hashes of the file at each position, and then computes pairs of files that have many chunk-hashes in common. Heuristics and human input are used to reduce "false positives" resulting from boilerplate or template content such as copyright notices placed at the start of source code files. Exactly duplicated files are reported separately to aid the human analyst. This allows a "cleaning effort [that] would be unthinkable without this data mining technology" and delivers significant business value (Forman et al., 2005, p. 399).

Prior work in conferences such as *CHI* and *CSCW* and journals such as *TIST* has featured advanced technologies and visualizations supporting text analysis and better navigation. Many of these assume without testing that the document comparisons made by the incorporated algorithms are a good match for human perceptions. That body of work is missing a rigorous test of how the algorithmic similarity measures being used match up with human perceptions. Study 2 below helps fill that gap by providing such a validation method in the task context of hyperlinks for navigation among topically similar documents, and applying it to a commonly used text similarity algorithm, the cosine similarity of LDA representations of the original texts.

Tools to enhance similarity and expose related content also generally incorporate a host of other design decisions that may impact their usefulness toward the stated goal. However, it is difficult to rigorously evaluate these tools against human perceptions of what the analysis is supposed to measure or support, or how these design decisions affect the way people perceive the content. The work below helps fill that gap.

0.5: Background: Overview

Online deliberation carries great promise for increasing participation in policy decisions, fostering interaction among diverse stakeholders, and increasing appreciation and tolerance for multiple points of view. Current tools are inadequate, either imposing rigid structures or leaving large collections of documents with insufficient structure to be understood or to be navigable. Machine learning techniques have been developed that provide powerful tools for addressing these problems of inducing topic structure, and there has been rapid progress recently on understanding how to attract committed participants to online communities. This thesis builds on existing knowledge with improved understanding of the needs, objectives, and motivations of users in a variety of roles as well as experimental evaluation of algorithms and the effects of exposing discussion related to peer-produced content on platforms for large-scale collaboration. To help answer empirical questions that can inform system design and advance relevant theoretical work, the program of research in this thesis spans a range from qualitative work such as an interview study to more quantitative controlled experiments.

Study 1: Interview Study of Platform Participants

Study 1 is a qualitative study examining participants in a live online deliberation platform, to help ground an understanding of participants' goals on platforms for public collaboration around aspects of a complex challenge, and barriers participants face in reaching those goals. It used a primary methodology of semi-structured interviews, described in detail below, supplemented by insights from published materials, participant-observation in the "Crowds and Climate" conference, and surveys and interviews conducted by other researchers.

1.1: Study 1: Introduction

Crowdsourcing, a relatively new term for "outsourcing a job once performed by a designated agent to an undefined, generally large group of people through the form of an open call, usually over the Internet" (Safire, 2009), is an increasingly common and effective way of getting things done (Howe, 2008). Motivations for people to contribute are not all that well understood, though we do know that people are "driven to contribute by a complex web of motivations, including [noneconomic motivations such as] a desire to create something from which the larger community would benefit as well as the sheer joy of practicing a craft at which they excel" (Howe, 2008, p. 15).

Even just within the set of paid crowdsourcing systems (i.e. those trying to use the Money motivation "gene" for collective intelligence (Malone et al., 2010)), we now have a spectrum of crowdsourcing platforms, with a wide range of task and payout size. Near one end, such as Amazon's Mechanical Turk ("MTurk[2]") or CrowdFlower, "microtasking" platforms provide short (e.g. seconds to minutes), simple tasks with rewards on the order of pennies to dimes (US$). Platforms like Fiverr offer tasks that take slightly longer and pay a few dollars. Requesters on Zeerk, Elance/Upwork (formerly ODesk) and similar platforms pay higher amounts for generally longer tasks.

[2] Most of the platforms listed here without reference can be found at "<name>.com."

Towards the higher end of the task-size-and-payout scale, the open call shifts from a mostly work-for-hire model with a high probability of success and payout to a contest model, where individual submissions have a lower probability of receiving a higher payout, related to the lower probability that any individual crowd member will be able to come up with a good solution. As examples, CrowdSpring, 99Designs, DesignCrowd, and NamingForce and offer contests for business branding with payouts often on the order of hundreds to low thousands of dollars. TopCoder offers payouts on the order of a few thousand dollars for solutions to programming problems. Continuing up the spectrum, Innocentive, HeroX, and Challenge.gov offer contest-based challenges with greater technical difficulty for payouts typically on the order of thousands to tens of thousands of dollars. Challenge sets like the Clay Mathematics Institute's Millennium Problems and Netflix Prize offer(ed) million-dollar prizes for solutions to extremely difficult math problems. The X-Prize Foundation is generally in the range of another order of magnitude higher, at the upper end of the range of task difficulty and reward for crowdsourced solutions to open challenges.

The contest-based incentivized prize model more common towards the higher end of that spectrum is not new. The Longitude Prize, created June 8, 1714 by the British Parliament, issued a prize-incentivized open call for precise measurement of a ship's longitude, and even then quickly overwhelmed the board of judges with a very wide variety of proposals on many challenging topics, even well outside the scope of the original challenge (Sobel, 2005). The Orteig Prize, first set out in a letter of May 22, 1919, offered US$25,000 for the first nonstop flight between New York and Paris (Bak, 2011, p. 28), later won by Charles Lindbergh. The Sikorsky and Kremer prizes later continued to incentive advances in aviation.

Incentivized-prize contests can increase the amount of resources dedicated to a particular problem by leveraging and helping create economic resources many times the size of the prize (Pomerantz,

2006). Some argue that incentivized prize models are more effective at promoting scientific progress than scientific grants, even while acknowledging that grants remain crucial (Leonhardt, 2007).

Modern attempts to use this model in the US government expanded from DARPA Grand Challenges (DARPA, 2014) to hundreds of others, gathering innovative ideas to solve real problems. Tens of thousands of participants (Obama Administration, 2015) include those who might not have been considered by more traditional bureaucratic processes (Gustetic, 2015) or even considered the problem without the challenge (Dorgelo, 2014). Tens of millions of dollars in prizes have been awarded (Obama Administration, 2015). The Obama Administration's Second Open Government National Action Plan, in addition to committing to increasing the navigability of government websites, committed new incentive prizes and challenges on Challenge.gov, expanded use of crowdsourcing and citizen science, and an interagency group to develop a toolkit of best practices and guidance related to open innovation, specifically including incentive prizes, crowdsourcing, and citizen science (Sinai & Smith, 2013; The Open Government Partnership, 2013).

While externally recognized as a valuable innovation in government (Dorgelo, 2014), significant additional research is needed to figure out best practices and best ways to use the crowd of participants (Santoso, 2015). Even the past White House recognized that people are incentivized by money and other causes, but still lacked a good understanding of just what various people are incentivized by, what their experience is, or how different platform design choices might match or interact with those motivations (Santoso, 2015). They recognized a need to better understand how to create a platform or environment where people who are incentivized by different things, but who want to work together to solve a problem, can do so effectively (Santoso, 2015). This is an active area of inquiry, with e.g. a high-level all-day user-centered design workshop to develop user needs for the toolkit referenced above (Gustetic, 2015) and its own federal communities of practice (DigitalGov, 2015; Lewis, Dunbar, & Crusan, 2013).

Having that knowledge could be helpful more broadly in the growing number of contexts where open incentivized-prize models are being used. This study aims to contribute to that body of knowledge with qualitative inquiry and grounded analysis methods, to better enable user-centered design for contest-based collaboration platforms, as one subset (Malone et al., 2010) of platforms for large-scale online collaboration which could potentially be used to address complex problems.

1.2: Study 1: Domain

Contest-based incentivized prizes are an important form of crowdsourcing solutions to difficult problems. This thesis focuses on platforms for large-scale and even crowdsourced collaboration on complex issues, and it is clear from the spectrum of existing crowdsourcing platforms that the more complex the problem posed, the more likely contest structures are to be used. It is also a structure being suggested (e.g. Leonhardt, 2007) and used (e.g. Malone et al., 2014; The White House, 2012) in response to the complex issue of responding to climate change.

One of the first "golden carrot" programs (incentivized-prize contest model) implemented to address climate change issues in the US was a $30M prize for a "super efficient" refrigerator (and then, other appliances), sponsored by a consortium of electric companies (Eckert, 1995; Joshi, Parrish, & Bauman, 1993). The program attracted as many as 500 responses (Joshi et al., 1993; Penn, 1993), but strict requirements on manufacturing capacity effectively limited full participation to larger companies like winner Whirlpool. Despite the amount, the prize money was not a primary incentive for the winner to participate; the potential to increase market share by getting a new product out and market its advantages was much more important (Eckert, 1995, p. 18). "The competitive aspect of the process proved to be a driving force" for participants (Eckert, 1995, p. 20). Despite the competitive aspect, a primary and unanticipated positive result from this contest was the unprecedented collaboration between industry, government, environmental, and consumer groups, building agreements between parties that held diverse and sometimes conflicting interests (Eckert, 1995, p. 22). A "genome" for

designing content creation tasks in a collective intelligence system presents a choice between "contest" OR "collaboration," not both (Malone et al., 2010). However, as demonstrated by the super efficient refrigerator prize and many others (e.g. Malone et al., 2017), these two are not mutually exclusive. There is clearly a complex relationship between competition and collaboration (and even "coopetition" (Brandenburger & Nalebuff, 2011)) in these platforms; this study aims to understand that more deeply from the perspective of participants in various roles.

For issues that are complex enough to be beyond the understanding of any particular individual or small group, solutions can only be developed through the collaboration of many. The individual pieces of such a solution may be brought in through a contest model. Then, if a platform can provide support for connecting appropriate sets of pieces together, progressively larger subsets of the "big picture" might be collaboratively assembled.

Incentivized-prize contest models are often described as being an example of collaboration (e.g. Malone et al., 2014; Obama Administration, 2015) and collaboration can be a primary positive result of such contests even when unintended (Eckert, 1995). How does this work? What goals, intentions, and experiences do participants have? How do people want to collaborate, if at all? What challenges and barriers do they perceive? How do those vary among people who have different roles, and who may have different interests in the outcome? How do the concepts of competition, collaboration, connection, and community come together? This study aims to answer such questions through a set of semi-structured interviews with participants in various roles on one particular contest-based collaborative platform trying to address the complex challenges of climate change (Malone et al., 2014).

1.3: Study 1: Position in Framework

In 1963, Engelbart published a framework for research about increasing the possibility, speed, and quality of both comprehension of and solutions to complex problems. This framework proposed two primary goals: "(1) to find the factors that limit the effectiveness of the individual's basic information-

handling capabilities in meeting the various needs of society for problem solving in its most general sense; and (2) to develop new techniques, procedures, and systems that will better adapt these basic capabilities to the needs, problems, and progress of society" (Engelbart, 1963). The former includes an examination of how individuals achieve their present level of effectiveness with the expectation that this examination will reveal possibilities for improvement; the latter includes ways to assess the research (Engelbart, 1963). Study 1 aligns with part (1) of this framework. Research questions for study 1 ask, "What are participants' goals and tasks in the current system, how do they organize work and handle information now, and what factors limit individuals' effectiveness in solving the problems they set out to solve?"

1.4: Study 1: Method

1.4.7: Research Setting

The questions identified above can be studied in the setting of a platform for large-scale online collaboration attempting to address a complex challenge (e.g. climate change) using the incentivized-prize model but in a setting intentionally trying to encourage collaboration, and placing a high value on encouraging integration of ideas from different submitters. Such a setting can be found in the MIT Climate CoLab, which is the research setting for this study. This choice of single case study is *revelatory* (R. K. Yin, 2013, p. 52) in that the setting was created to help support research and exploration into practical large-scale computer-supported collective intelligence systems. This choice of single case study is also likely to be *representative* in that the roles and experiences identified here are likely to be common (in the sense of R. K. Yin, 2013, p. 52) to similar present and future platforms.

Inspired by examples of successful large-scale collaboration elsewhere, such as Wikipedia or FoldIt, the authors of "How Millions of People Can Help Solve Climate Change" (Malone et al., 2014) created the Climate CoLab to be a global platform for collaboratively developing and evaluating proposals for what to do about global climate change. In an annual series of contests, its members (approximately

87,640 registered as of the end of April 2017) have collectively produced, commented on, and voted on over 2000 proposals, each typically 1500-3000 words in length (Climate CoLab, n.d.-c), on a wide range of climate-change-related topics. As of the end of the 2014 contests, the site had over 30,000 registered users and about 1,000 proposals. Since 2013, several contests have been run in parallel, each addressing different subtopics (such as Transportation, Waste Management, Buildings, and Energy Supply) or calling for plans that integrate a number of other proposals (Malone et al., 2017). Although participants come from many countries, the interface and proposals are generally in English.

Each CoLab contest has one or more Advisors, invited experts who help develop contest materials, recruit judges, and connect winners with people who might help implement their proposals (Climate CoLab, n.d.-a, n.d.-f). Advisors help invite 4-5 expert Judges (often including themselves), who are asked to spend only a couple hours in a multi-round judging process, each time reviewing up to 10 submitted proposals (Climate CoLab, n.d.-g) plus one or two hours on conference calls to discuss proposals and select finalists and winners (Climate CoLab, n.d.-c). Each CoLab contest is facilitated by a few Fellows who volunteer a few hours per week to help coordinate the other stakeholders and shape the contest activity. Fellows work with contest Advisors to write the prompts, promote the contest and its winners, comment on proposals and pre-screen them before judging, and facilitate the logistics of judging such as conference calls and accumulating judges' comments (Climate CoLab, n.d.-b).

In written work, the organizers describe the Climate CoLab as "a place where real exploration and novel thinking occur" because of the diversity of members' educational and other backgrounds and deep interest in what to do about climate change (Climate CoLab, n.d.-c). Even on controversial issues, the platform explicitly welcomes "all positions" that can be supported via the scientific method (Climate CoLab, n.d.-d), and hopes to find "new angles, new perspectives, new ways of looking at things" (Larson, 2014). Its director claims, "We now have a new way of solving really big, hard, complicated problems at a scale, and with a degree of collaboration that was never possible before" (Daniel, 2014; Larson, 2014).

The stated primary goals of Climate CoLab contests are to:

- Harness the collective intelligence of large numbers of people around the world to create proposals for what humanity should do about global climate change (Climate CoLab, n.d.-e).
- Help to educate the general public about the real issues involved in global climate change (Climate CoLab, n.d.-e) and change the public dialog on climate (Daniel, 2014).
- Provide a sense of optimism to help people see there is something they can do about climate change (Daniel, 2014).
- Provide a large-scale test of new collective intelligence approaches, building on examples like Wikipedia and Linux (Climate CoLab, n.d.-e).
- Provide a neutral forum where the best ideas and information about global climate change can be shared (Climate CoLab, n.d.-e).
- Help launch ideas on a bigger scale (Daniel, 2014).

The number of registered members as of January 2015 appears to be about 12% of the number of people who have visited the site (Duhaime, Olson, & Malone, 2015a); the distribution of participation at various levels seems consistent with the power law and "1:10:89 rule" observed elsewhere (e.g. Howe, 2008, p. 227). In July 2014 and on subsequent occasions, a survey was e-mailed to registered members for the first comprehensive characterization of the MIT Climate CoLab community, including an understanding of the impacts CoLab participation has had on participants, and to help better understand participant motivations for joining. A little under 10% responded to the survey, with demographics analyzed and reported on by other members of the Climate CoLab research team (Duhaime, Olson, & Malone, 2015b). Participants reported that since joining the site, their understanding of climate change, perception of the issue's importance, and level of activity to address the issue all generally increased (Duhaime et al., 2015b). The survey revealed that even individuals who were not previously involved in climate-change related activities were successfully able to contribute

innovative, valuable ideas about how to address the challenges, and that while factors such as experience, gender, and nationality (non-US) did impact members' probability of submitting a proposal, they did not impact the likelihood of a proposal being selected as a finalist (Duhaime et al., 2015b). Approximately 60% of survey respondents indicated that they would be willing to be interviewed, and provided an e-mail address for that purpose. Most interviewees in study 1, other than organizers, were chosen from this list.

1.4.8: Overview of Interviews

The primary method for study 1 is semi-structured interviews, as described in (Corbin & Strauss, 2014, p. 39) with Climate CoLab participants in various roles. A diversity of roles was sought expecting that interviewees with different roles may have different goals, expectations, and perceptions of what might help or hinder them in their tasks. In each case, I hoped to better understand the interviewees' goals, hopes, and motivations for participation. What do they hope will result from it? What did they do in pursuit of those goals? What tasks are they trying to accomplish and what strategies are they using to accomplish those? How well do those strategies work? What barriers do they perceive are limiting their capabilities? What factors impeded or facilitated their abilities to reach their goals? What do they expect the platform to support? Who do they expect will be affected by what they do? What are the collaboration and coordination problems their participation, hopes, and goals give rise to? From their experience on the existing platform, what do they think best supports their goals, and what do they perceive is lacking? Results can deepen our understanding and guide future work for developing such platforms.

In selecting participants, I sought out people who are in roles doing tasks that are likely to be common to platforms for large-scale collaboration on complex issues, especially those that are contest-based or using the incentivized prize model. One common role is that of platform organizer. I interviewed platform organizers to understand their vision, the guiding principles behind some of the

relevant design choices, and their perception of the current state. A second task likely to be common to such platforms is that of comprehensively reading some subset of content, whether for quality control, community management, potential implementation, or some other reason. Within the Climate CoLab, this task is part of the roles of Catalysts and those involved in the judging process. Interviews with these individuals help increase understanding of how they organize proposals for their judging or reading through the contributed content. Platforms of interest for this study also all need a role for content authors, as well as readers who may be not contributing proposals themselves. I interviewed people in both of these roles as well.

A hypothesis from the literature described in the background sections of this thesis is that automatically structuring content along topic lines seems promising as a solution to some of the problems participants may be experiencing. To explore this area further, some interview questions focus particularly on navigability to identify whether or not this is an area that interviewees think is important and/or in need of improvement, and if so how they think it might be done to best support what they are trying to accomplish. Where appropriate, participants are specifically asked about what value (or negative consequence) they might perceive in the addition of connections between related proposals, so that the results of Study 1 can be used to add qualitative contextual interpretation to the results of Study 2, for guiding future work in this area.

1.4.9: Data Collection

I actively observed the 2014-16 contests as they ran, read relevant background literature including materials on the CoLab website, was a participant-observer in one contest, reaching Finalist status (Towne, 2014), and attended the 2014-16 "Crowds and Climate" conference to become familiar with the Climate CoLab and provide a background of knowledge for the data collection and analysis in this study. Between the end of the 2014 conference and the beginning of the 2015 contests, I conducted 25 interviews with 3 organizers, 2 catalysts (whose role it is to read through and comment on all proposals

in one or more contests), 5 individuals formally involved in the judging process (as fellows, judges, and/or advisors), and 14 participants, half of whom were proposal authors (covering the four award statuses present in most contests). Quotes below from interviewees are denoted with an "O," "C," "F," or "P" respectively, each with an arbitrary but consistent number for identification in this study. Non-interviewees referred to also received codes like these, so the range of numbers used in quotes exceeds the number of people interviewed. Some participants (in the last category) did not actively participate in the year preceding the interviews; this was an intentional selection to help increase the diversity of perspectives and range of recent participation. Interviewees came from four continents. Interviews typically lasted just under an hour. Interviews were conducted over the Internet, most commonly via videoconferencing to help reduce crosstalk, and audio-recorded with participants' express consent. One interview was text-only, limited by power and bandwidth restrictions at the interviewee's location. The size and diversity of the set of participants is similar to an interview study by Dabbish, Stuart, Tsay, & Herbsleb (2012), with participants on a website enabling large-scale collaboration around software development.

Before requesting an interview, a participant's tracked activities (as displayed on their Climate CoLab member page) and role in the Climate CoLab were reviewed. Potential interviewees received a personalized interview request, and a single follow-up if they did not initially respond. A majority responded and were interviewed. One potential interviewee indicated plans to participate but stopped responding during scheduling discussions, for an unknown reason. The semi-structured interviews were guided by a list of prepared topics and potential probes (questions) for exploring those topics. The semi-structured format permitted exploration in greater depth on particularly interesting concepts or themes that interviewees raised, dynamic re-ordering of topics to better fit the flow of conversation, and omission or modification of certain questions for time or relevance purposes. Live notes were taken

during most interviews for the primary purpose of being able to return to earlier topics with follow-up questions without interrupting the speaker in a point the participant was trying to make.

All interviews were transcribed, most of them with the help of a professional transcription firm on the first pass. All transcripts were reviewed and corrected for accuracy, including those that had been professionally transcribed. The average corrected transcript was 7,783 words (21 pages) in length.

1.4.10: Data Analysis

I analyzed the data using grounded methods, guided by (Corbin & Strauss, 2014). Following transcript correction, transcripts were partially anonymized by replacing most direct identifiers with a code, indicating the role of an individual or the general point(s) of certain other specifically identifying statements. For example, references to the participant's geographical location, organization, or proposal were replaced with less specific indicators noting this is what was referred to. When the same content was replaced by an anonymization code multiple times, the same code was used to permit a reader to identify that the referenced content was the same in those different places. This was done under the guidance of Corbin & Strauss who point out that it is the meaning given to events in the interviewee's context, rather than the event itself, that is of primary interest for study (Corbin & Strauss, 2014, p. 25). This step replaced specific uniquely-identifying events with a summary of meaning that abstracted away less important details and made it easier to observe patterns among what the interviewees were saying. This partial anonymization also helped from an IRB perspective in partially protecting specific identities, at least from a casual reader. The semi-anonymized transcripts were then uploaded to Dedoose, a web application facilitating qualitative research and grounded open coding methods (see e.g. Corbin & Strauss, 2014).

During the processes of transcript correction and partial anonymization, several points that seemed especially relevant to the research goals were flagged for further attention. After the transcripts were uploaded to Dedoose, these points were explicitly sought out and coded on Dedoose in an open coding

process, in which over 1200 excerpts were coded. In many cases, other relevant points not expressly flagged in the offline transcript (but perhaps nearby in the text) were also coded. Descriptive codes were developed as seemed appropriate, with a bias towards more specific and descriptive codes that would make it easier to later find particular points that interviewees had made.

Those codes were grouped and categorized in a multi-level tree. Content included quotes related to goals, motivations, difficulties or barriers encountered, perceptions of community, perceptions of the judging process, competition and collaboration and the relationship between them, diversity and/or potential for synergy among the proposals, proposal quality, the value and use of comments, the value and use of the Related Proposals section in current proposals, the way people currently navigate through the site, alternative or potential future ways of navigating through the site, interviewees' use of social media especially as it relates to this platform, and interviewees' intentions around future involvement. Each of those categories had several more specific subcodes identifying more specific meanings within the theme, producing a structure of codes helping identify patterns within the data and connect thematically similar quotes across interviews.

1.5: Study 1: Results

Goals and motivations seemed to fall into three main topic areas, as described in this section. One is the collective intelligence or crowdsourcing vision of the site, a second is the vision of addressing the challenge posed by climate change, and a third (especially for proposers) is potential funding and/or recognition.

1.5.1: Primary Motivations

1.5.1.1: Collective Intelligence

Organizers expressed that one of the main goals behind the creation of the Climate CoLab was to explore techniques and platform design decisions for collective intelligence.[3] Organizers have a strong interest in supporting research on collective intelligence towards solving complex problems, more generally than e.g. an organization focused on a particular field of content might be. Interviewees in other roles also indicated that they were attracted to the platform because of the vision around collective intelligence and large-scale collaboration (e.g. P29: "*the collective intelligence idea caught my attention. That is why I'm very interested in Climate CoLab,*" or P5: "*people, they assumed I was there for climate change. And, of course, I am, indeed, very interested in climate change. But I'm there really for collective intelligence.*"). This indicates that these or similar people might also be participants in other such platforms and projects that do not yet exist, and so understanding their motivations, goals, and barriers on this platform would usefully contribute to the scientific literature read by those who might want to build similar platforms in the future (or improve on existing ones, including but not limited to the Climate CoLab). Interviewees' definitions of collective intelligence included permitting everybody to contribute (P8: "*if everyone can contribute, then that's how – it's collective intelligence*") in "*small portions at a time*" (also P8) "*to do very large scale collective problem solving*" (O2).

1.5.1.2: Addressing the Issue of Climate Change

The most common motivation heard from interviewees in all roles was, unsurprisingly, a desire to address issues associated with climate change. As F9 put it, "*I didn't do this just for fun. I'm doing this because I want to see change.*" Interviewees generally believed that anthropogenic climate change was an important issue which required response. They generally assumed that a consensus exists about the

[3] The platform was developed and continues to be maintained by the MIT Center for Collective Intelligence. Subsequent to these interviews, MIT Solve (http://solve.mit.edu) also adopted it for use in other content domains.

problem but not the appropriate response(s). People wanted to see impact resulting from the work that was happening on the site. People wanted to spark discussion, change perspectives, change policies, implement solutions, and leave a better legacy for future generations. They wanted to see what impact was resulting (e.g. P32: *"I'd really love to see it get to the point where there's like, the done-it list. You know, a whole list of like, "Yeah, we did it." You know, like went from an idea, actually got the technologies developed, actually got it implemented with the government; here's the photos of it happening."*). If they didn't think there was much impact (e.g. P22: "*weatherizing my house probably has a greater effect on the world than, like, generating good ideas on this website*"), they expressed dissatisfaction (e.g. P12: *"nothing comes of it, I have to say. It's a very disappointing, frustrating, process. I mean, it seems like a waste of time"*).

Participants expressed a lack of knowledge about who if anybody was reading proposals, speculating that it might be potential policy implementers, leaders of various sorts, people with a lot of resources who fund or back initiatives like these, technical experts (i.e. scientists and engineers) on topics related to climate change, authors of other (especially similar and/or competing) proposals, professionals in the community of practice around climate change responses, and potential collaborators. However, they acknowledged that there were barriers for the intended audience (e.g. P5: "*there are a lot of proposals there for them to look at*").

1.5.1.3: Funding and Recognition

Participants, especially proposers, mentioned the possibility of funding as a motivating factor in their participation. They mentioned not just the US$10,000 grand prize, which would be awarded to a single winning proposal, but also the possibility of an investor or implementer learning of their idea through the Climate CoLab and choosing to fund further work on their project, or help provide other important resources. A number of interviewees were motivated by the networking opportunities that participation presented. Some participants thought the $10,000 grand prize led to too much

competition and reduced willingness and incentives for collaboration. Some participants were motivated by the potential for recognition, perceiving that an award status bestowed on their idea gave it some external stamp of credibility from a well-respected institution (MIT). Proposers sometimes used the online presentation of their proposal as a handy link to refer people to when talking about the idea that they had proposed, well after the end of the contest. Some participants were motivated to participate simply because they enjoyed the stimulating challenge (P21: "*and there's a real – there's an enjoyment in taking on a big challenge*"). Others saw ways that participation in the Climate CoLab could support their own aligned work or goals outside the Climate CoLab.

1.5.2: Difficulties in Reviewing Proposals

1.5.2.1: Volume

Interviews with people in the Catalyst, Fellow, and Judging roles also indicated that volume of proposals was a challenge (e.g. C5: "*I was not expecting so many projects. I was a little bit shocked about the number…reading all that was absolutely impossible*") that limited the quality of possible results (e.g. F1: "*fewer proposals would have been better and given the judges more time to kind of offer more substantive feedback.*"). Interviewees described how the current judging process would only scale up linearly or worse, though organizers expressed a desire to see a great deal more proposals. In 2014, the contests had an average of 28 proposals and a maximum of 53.

If the effort required for the judging process scales worse than linearly, it may because of coordination difficulties among the judging teams, which interviewees reported as a challenge in the 2014 contests. Much of this was simply the logistics of trying to find judges and schedule conference calls across time zones around the world, some lack of clarity about roles, and some unreliability or unexpected life events affecting individual volunteers. Many, but not all, interviewees felt that the judges were overloaded in terms of the time required to do a great job in the judging process, including both volunteers involved in the judging process (though they did spend the time) and participants (e.g.

P16: "*we got the feeling that maybe the judges were being pressed for time, so they didn't really get to really read the proposals*"). On a related note, participants also felt that preparing proposals took a lot of time and effort. P16 reported, "*I personally probably put in 100 hours… But I think <two team members> put in 9 months of probably at 40 hours a week… They worked -- And you know, a lot of that was editing and making sure the links worked, but it was like birthing a baby for them.*"

One organizer suggested that if they were successful in attracting a higher number of quality contributions, "*we could find more judges, that's a potential, or we could ask judges to volunteer to take on more*" (O1). The former of those options increases coordination requirements; the latter increases individual burden and may restrict the pool of potential contributors in this role, as discussed in (Towne & Herbsleb, 2012). For the Climate CoLab or a similar platform to run effective large-scale contests, figuring out how to handle a large volume of proposals is an important challenge.

1.5.2.2: Quality

Proposal quality was an issue reported by interviewees who were reading through reasonably large sets of proposals, especially if they did not have award status indicators to guide them. Most interviewees believed that the set they were reviewing contained some high-quality proposals, but finding that signal in the noise was a challenge. People involved in the judging process used an assessment of "how well thought through is this idea" as a first-pass consideration, especially for screening, which led to a set of winners with small and focused but well-developed ideas. Some interviewees, including platform organizers, saw this a problem and wanted to see some big new ideas that might not be thoroughly developed in detail, but which might have potentially greater impact on addressing the challenges. Many interviewees mentioned that the proposals they saw on the Climate CoLab duplicated ideas they had heard about elsewhere, without connections to those external ideas.

1.5.3: Navigation

1.5.3.1: Current Strategies

The current dominant method of navigation on the site is for a user to arrive at the Contests page and see the list of specific topics, find the one(s) of greatest interest, and click in to the list of proposals in that contest. Many interviewees expressed that the contest structure was a reasonably good way of grouping topically similar proposals and guiding a reader's navigation, though there is some overlap which leads to miscategorization of proposals and missed connections between proposals in different contests. The contest grouping was a great fit for proposal authors who read other proposals primarily to learn more about who they were competing against. There was little mention or use of the taxonomy of ideas which is already built in to the Climate CoLab (more obvious in Outline View).

After clicking into a contest, viewers then read down the list of proposals, scanning the titles and summaries (<280 characters) in that list to decide which ones they wanted to click through to. Some interviewees found this difficult because especially the titles are sometimes more creative than descriptive, not helping the potential reader understand what they would find clicking through to a proposal. Some readers considered the award status of a proposal (which also moves it up to near the top of the list) in deciding what proposal to review or click through to.

In the top navigation bar on most Climate CoLab pages, there is a "Search" button which opens a drop-down text field for entering a query. Some participants mentioned using the keyword search functionality, but expressed a perception that it does not work very well. Even when a searcher knew of a particular proposal that had been referenced (e.g. a winning proposal around fast breeder reactors for nuclear power), it could still be hard to find in this particular keyword search. Some of these difficulties may simply be engineering challenges, but at least some are inherent to keyword search (e.g. C5: "*since you don't know what you're searching for*" or P31: "*if you don't know the exact words*").

Several interviewees noted that there was room for improvement regarding navigability, even though a partially overlapping set noted that the contest structure was reasonably good at organizing the proposals. For some interviewees, but not all, navigability posed a barrier to what they were trying to accomplish on the site, whether reading (P31: "*the first time I spent, I may have spent upwards of a half an hour on the site digging around and then after that probably a little bit less time, feeling like I was getting sort of swamped…I didn't successfully find what I was looking for*") or writing a proposal with an idea "*that I've thought about submitting many times, and I'm like, I don't even type it out. Because I can't figure out where it would fit, how to position it.*" (P32). P31 was a reader who has resources for startups, a persona that many (but not all) proposal authors were trying to attract; those authors' motivation to contribute was based in part on hoping readers like P31 would find their proposal and help support it. P32 was a potential proposal author who is outside the mainstream scientific dialogue on climate change, but even from this "outsider's" position has had past successes that help address the issue and during the course of an extended interview described three potential proposals, each of which could have fit in well with the overall set of proposals on the Climate CoLab. P32 was of a persona the organizers wanted to see more proposals from.

Each proposal has a "Related Proposals" field that proposal authors can manually fill out with whatever text and/or links they choose, but was commonly left blank. Anecdotally, it appears that proposals reaching award status were more likely to have made good use of this section. F1 speculated that this might be because "*the people with the strongest proposals actually took the time to fill out the proposal completely including the related proposals section.*" This section prompted some proposal authors to look for related proposals, as intended by the organizers. In at least one case, an interviewee left a comment on one of the related proposals he found, and was subsequently asked to join the team as a collaborator on that proposal. P16 suggested "*Maybe there should be a requirement that you find another proposal that can work with yours or something like that,*" reflecting on the difficulty of that

task but believing it likely possible given some creativity and the volume of available proposals. Interviewees also described potential value in curating and limiting the number of links because the marginal value of each successive link decreases, especially once there are already 10-12.

1.5.3.2: Potential Improvements

When discussing potential opportunities to improve navigability, participants came up with a range of ideas of how they think it would work. Many suggested keyword tagging, sometimes with qualifiers (like the ability to merge synonymous or highly related tags). Some participants discussed having a "*grand topic map that tries to keep everybody's ideas roadmapped*" (P20) or asserted that "*If you don't carry out a hierarchical analysis you cannot – you will not be able to figure out how to put proposals together*" (P29) though the scalability of this manual approach is questionable, and a taxonomy already exists on the site.

Many interviewees indicated that there was a lot of potential for synergy between proposals, though sometimes hard to find. The discovery of these synergies seemed to take place more at the in-person conference following the contests than online.

Many interviewees said that drawing connections between related proposals would be useful to them. For some, it was expected (P16: "*Wasn't there going to be some kind of follow-up with that? … I'm thinking that I remember hearing or reading about something where some proposals – or people who'd written proposals would be connected if they had similar proposals or encouraged to work together. I don't know if that was – isn't that part of the going through the rounds of judging?*"). For others, it would be motivating. If connections between proposals helped people who shared a proposer's interest and idea find it, and the proposer found out (e.g. through direct communication), P12 would feel "*on top of the world. That would be great.*" P21 thinks there is a "huge opportunity" for synergy among proposals and that "*it's a real missed opportunity if that isn't taken advantage of or if it doesn't become more of a point of focus – I mean, if I were in an administrative capacity with respect to*

the Climate CoLab, that would be a pretty high priority for me. Exactly how to implement that would be an interesting subject for discussion, you know?"

When asked "*If you had a magic visualization that made just the things you were looking for very clear and easy to see, what would those things be?*" P5 responded "*it would be a person who was interested in my project who could just go – so if, indeed, I received a comment from someone unknown who just goes, "Hey, I love your project, and whatever, but I don't know if you know about this other thing that's already happening, which is such and such," where I could kind of go, "Oh, you're right. That is a place where we could really hammer out some synergy really quickly." You know, that would be the thing that in my case would have been the magic … happening. And so it's gotta be easy to – not easy, but it's possible to think about that happening in this wonderful connected age where we have an infinite information resource and Google-based algorithms and that sort of thing. But that would be it.*" Elaborating, P5 referred (hypothetically) to an algorithm that matched keywords about a proposal and the area of effort it sits in "*and just goes, you know, 'Here is a project that has a 68 percent match, and that actually is pretty high considering the relative scarcity of your keywords,' or you know, whatever, then, yeah, I would click through to that.*"

When asked very generally about possible improvements to the CoLab, P12 said "*I think one thing you could do very easily is group us together. And if there's somebody else working on <functionality provided by P12's proposal>, put us in touch with each other. Like … kind of group them, in problem, technologies, or whatever … because if we're working on an idea, we don't have time to have time to sit there and peruse through a bunch of random ideas. So I don't know if you're codifying and categorizing, you could then redistribute them to the entrants. Who knows what you might inseminate in that process? [Those connections should be in the form of a] link. A link is simple. A link to their entry. You could have a little email goes out, says, well, 'Thank you for your entry to Climate Co-Lab, and here are some entrants in similar categories you might enjoy.' Stick a bunch of links in… those are filtered out for*

you already. You might say, that makes my idea work better, and you could get in touch with each other. And they come back the next year with an even better entry."

One interviewee suggested that especially if there was more duplication from a greater volume of proposals, "*You could probably do text analysis of the text, and if there's lots of similarity, you could probably say hey, guys, this looks pretty, you know – this looks like there's a lot of words that are the same here. You all ought to maybe check each other's work out… build a network map of how close or distant the proposals are based on their language.*"

1.6: Study 1: Discussion

While this interview study examines participants on only one particular platform for large-scale collaboration addressing one complex issue (climate change), the factors underlying motivations and barriers to participation in such efforts are likely to generalize to other platforms. Some details may change from the specific platform to more general conclusions (e.g. a motivation of "caring about addressing the issue of climate change" might be generalized to "caring about addressing the issue,"), but many of the general challenges, roles, and methods of work are likely to be the same for large-scale collaboration projects addressing other complex challenges. Questions around means and methods for navigability, identified in (Towne & Herbsleb, 2012) as a relatively general challenge, are also not specific to a given platform at a conceptual level. The vision around collective intelligence and large-scale collaboration was also one of interviewees' primary motivations to participate, indicating that our participants may help reveal perspectives of individuals who might be likely to participate on such platforms more generally.

The interview results concerning navigability indicate that increasing navigability would help various participants be better able to achieve their goals on the site. Study 2 presents methods for testing an automatic method of detecting similarity between documents, as part of a potential solution to this issue, somewhat aligning with the second part of Engelbart's (1963) framework cited above. Reaching a

more complete solution requires a better understanding of human perceptions of similarity (e.g. is it a single construct or multi-dimensional?), a way to measure whether or not an algorithmic similarity measure does indeed match human perceptions of similarity, and an assessment of how well the current state-of-the-art method used in these kinds of applications does or does not match human perceptions of similarity. Study 2 addresses these questions.

Study 2: Evaluating a State-of-the-Art Topic-Modeling Based Similarity Measure

2.1: Study 2: Chapter Summary

Most theoretical approaches to bibliographic classification bring together concepts, as contained in documents, based on similarity, although this is challenging because there might be many relatively orthogonal dimensions in which concepts can be similar or different (Mai, 2010). Since the relevant design questions focus on fundamental ways of structuring the deliberation space, experiments are also needed to gauge the effectiveness of different measures of similarity between documents. This study contributes a method for validating such algorithms against human perceptions of similarity, especially applicable to contexts where the algorithm is intended to support navigability between similar documents via dynamically generated hyperlinks. Such validation enables researchers to ground their methods in context of intended use instead of relying on assumptions of fit.

Experiments in this study measured how a state-of-the-art similarity algorithm (cosine similarity of a Latent Dirichlet Allocation topic vector) captured human perceptions of similarity. Participants were presented with three randomly-ordered ideas from a subset of the 2012 SAVE Award dataset. One was designated (for selection and analysis, but not to participants) as a focal document. Participants were also presented with the "most similar" document according to this relatively simple algorithm, and a document that was not similar according to this metric, either with the same most-probable topic (≈50% of experimental runs) or not (≈50%). Participants were asked to choose the most similar pair and explain why, qualitatively.

Two thirds to three quarters of the time, participants chose the same pair as the LDA-based similarity metric as being the most similar. The remainder were about evenly split between pairing the "distant" document with the "focal" or the "near" document. There were no significant differences in the demographic questions between people who chose the same pair as LDA and those who did not. About 12% of participants' selections for which pair was the most similar could be explained by a

participant only understanding one or two of the ideas, as measured by a Likert item response and as suggested by the qualitative data. About 21% of participants' selections seemed to be based on participants' assessments of idea quality, measured similarly. These "explanations" were each more common in cases where participants disagreed with the LDA-based measure; combined they could explain 36% of such cases.

A closely related experiment gathered data about the human-perceived similarity of pairs using a new, highly reliable seven-item scale, which correlates with LDA-based similarity. A third experiment gathered data on the coherence of salient topics and the perceived fit between the documents and the lead topics they were associated with in the experiments. In general, agreement between human and algorithmic measures of similarity was higher when the documents "fit" the dominant topics well. Further, human and algorithmic measures of that "fit" were correlated ($r=.506$ $p<.0005$, $n=65$), suggesting that some algorithmic self-evaluation of confidence might be possible.

2.2: Study 2: Introduction

As discussed in section 0.3.3 above, ideation tools can capture innovative ideas from large, diverse populations, with potential to generate billions of dollars in value and address large-scale urgent problems such as disaster relief. Organizing these contributions so readers can navigate through them is a challenge. In many such tools, including that which collected data used in this study, ideas are often organized by post time or upvote count, which are problematic for the reasons described above. Algorithmic help for navigating along conceptual adjacencies can be useful (Johnson, 2010), as long as human readers perceive conceptual connection in those links. Other work assumes that users have well-defined views of similarity for the purposes of organization, and assumes that the computer can learn these (e.g. Huang & Mitchell, 2007).

As discussed in section 0.4.6 above, and expanded on in more detail in the next section, many algorithm-based technologies and visualizations designed to do text analysis and support navigation

assume without testing that the document comparisons made by the incorporated algorithms are a good match for human perceptions. In many cases, papers about such tools neither focus on specific tasks (describing a general intention to support exploratory analyses) nor evaluate tools with tasks that have measurable, comparable outcomes. This match should be empirically validated for the intended application and tested against reliable human judgments of similarity, with differences investigated and understood.

Study 2 contributes to this body of work by presenting a method for measuring how well an algorithmic measure correlates with the links that humans would make, and under what conditions it performs more or less well. The method is used to evaluate a widely used similarity measure, specifically cosine similarity based on Latent Dirichlet Allocation (LDA) topic vectors. Study 2 presents an application of this analysis and caveats to consider when using the measure. It also provides methodology validation measures (e.g. inter-rater reliability), so that other researchers might be able to more easily evaluate proposed advances over this state-of-the-art analysis technique. As graphically illustrated in Figure 1, Study 2 also analyzes the topics, documents, and reasons where human similarity judgments were most different from the LDA-based cosine similarity measure in order to understand the reasons for divergence and provide insight into particular strengths and weaknesses of using this algorithm in this application. This evaluation is useful for others who may wish to use similar algorithms in their own work, as well as those who wish to develop and test improvements to those algorithms, especially as they apply to increasing navigability.

2.3: Study 2: Literature on Evaluating Similarity Measures

As a direct example of using the cosine similarity of LDA topic vectors to compute text similarity and support navigation based on topical similarity, and a system whose utility depends on the match between LDA similarity and human similarity judgments, the Stanford Dissertation Browser relies on this measure to help users explore over 9,000 thesis abstracts. The paper describing that work cited one of

the field's shortcomings as a lack of validation mechanism or external ground truth to assess similarity measures (Chuang, Ramage, Manning, & Heer, 2012, p. 447). Boyack et al. (2011) use LDA and other measures to compute pairwise similarity among 2.1 million biomedical publications (medical subject headings, titles, and abstracts), and acknowledged the value of human-based validation measures, operationalized there as connections through grant acknowledgments.

TopicNets also uses LDA to support exploration of a document corpus, by providing an interactive graphical environment showing LDA topics which the documents are connected to and through. Evaluation of the human usefulness of the system is, in that paper, done by the authors' generation of graphs, interacted with and explored by an individual with expert knowledge of the dataset (including authors) (Gretarsson et al., 2012).

Some systems use LDA to identify latent communities on platforms where users interact via text, and display the relationships among items based on these latent topic-based communities. For example, a system called Pharos uses LDA to model latent topics of text content and clusters documents based on their most-probable topic, illustrating communities and changes in them over time. In that work, evaluation consisted of ten users from IBM completing two identification tasks (of top authors & blog posts on a specific topic) with and without the tool being available to evaluate its usefulness (Zhao et al., 2011). Yin, Cao, Gu, & Han (2012) somewhat separate the concepts of topic and community, modeling community-based latent topics in user-generated content on social media sites. This method is evaluated by the authors' qualitative impressions and comparison to the results of other methods. Work by Zhang, Qiu, Giles, Foley, & Yen (2007) discovers communities by modifying LDA for application to social graphs, testing the new method in research co-authorship networks and successfully identifying groupings of researchers within the same institution or research area. Introne & Drescher (2013) note that topic modeling techniques do not support analysis of multi-party dialog well, in part because of the

greater dynamism of dialog, and design an extension of its concepts to model communities of words over sequences of replies.

Some studies have examined the match between algorithmic and human perceptions of similarity using a stand-in such as a human-built hierarchical Open Directory instead of direct evaluation (e.g. Haveliwala, Gionis, Klein, & Indyk, 2002). Others have used the DARPA[4]-organized TREC[5] collections with an abstracted retrieval task ranking document sets as more or less relevant (according to human judges at NIST[6]) to given information-need statements (Voorhees, 2007) rather than to each other. Mani (2001) presents several methods for evaluating the match between a long document and summary of it, or between two summaries. Dinakar et al. (2012) state that an approach using LDA performed better than TF-IDF in a story matching task where participants completed several two-item pairwise similarity evaluations.

At the individual word level, Faruqui et al. (2015) computed the cosine similarity between feature vectors that incorporated information from lexicons and large corpora, and compared these algorithms against benchmark data sets containing pairs of English words that had been assigned similarity ratings by humans. The Spearman correlation of these two similarity scores (human and algorithmic) was used as the algorithms' primary evaluation measure, and feature vectors with a higher correlation were said to produce better results than those with a lower correlation. The interquartile range of Spearman correlations reported in that paper's Tables 2-4 is (.580, .737).

Sizov (2012) extended LDA to estimate resource similarity based on both tags and geospatial data in support of automatically organizing, filtering, and recommending content on social media. A panel of

[4] Defense Advanced Research Projects Agency (USA)
[5] Text REtrieval Conference
[6] National Institute of Standards and Technology (USA)

five computer scientists working in the field of social media research found the new system to support those tasks well (Sizov, 2012). Other work in *ACM TIST* has applied LDA to mobile phone location/activity data to find patterns (Farrahi & Gatica-Perez, 2011) or clusters of sociological interest (Joseph, Carley, & Hong, 2014); still other work in the same venue has focused on speeding up parallel computation of LDA models (Z. Liu, Zhang, Chang, & Sun, 2011).

LDA is used widely, but assumptions about the extent to which the meaning attributed to topics is consistently represented in the texts the model assigns the topic to are often unchecked. This motivates a need for new quantitative methods for measuring semantic meaning in inferred topics, based on human perception, as it is a quality not well measured by traditional algorithmic metrics (Chang, Boyd-Graber, Gerrish, Wang, & Blei, 2009).

The work closest to Study 2 in task and goals is Lee et al.'s (2005) investigation where students rated pairs of short news stories on a 5-point relatedness scale. The authors report a human-rater-agreement level of .605, computed as the mean correlation between each rating and the others for that pair. They evaluate a number of similarity algorithms and find that the best simple models have a correlation of about 0.5 with human judgments. Their most complex comparison is cosine similarity based on Latent Semantic Analysis (L**S**A), trained on a 364-document corpus. The best L**S**A models correlated about 0.6 with human judgments. In general, they find that the best models can detect only a subset of the highly similar document pairs, which suggests a need for alternative, more nuanced, models of text document similarity. This work measures human perception of similarity in more detail and with a larger, more diverse group of human evaluators, and evaluates an L**D**A-based cosine similarity measure against human judgment. It generally finds that the more advanced model seems to offer a closer match to human-perceived similarity, while identifying shortcomings of the L**D**A-based measure which could be considered in motivating further improvements.

2.4: Study 2: Experiments: Overview

Study 2 is a series of experiments designed to explore the connection between LDA and human conceptions of similarity, as illustrated in Figure 1. Following an overview here, section 2.5 describes models and materials selection and sections 2.6 – 2.10 provide more detail about each experiment.

I begin with a task that models selection of links between ideas, as might be used to help address some of the challenges discussed in the background sections above: given three documents, I ask people which two are most similar and to explain why in free text. I construct the sets of three in ways that help assess the level of agreement between participants and the algorithm, but more importantly, explore the sources of failure of algorithmic similarity to match human judgment. I construct sets of three documents that always contain a second document that is highly algorithmically similar to the first, and a third that the algorithm ranks as rather different from the first. Those three documents are randomly sequenced for presentation.

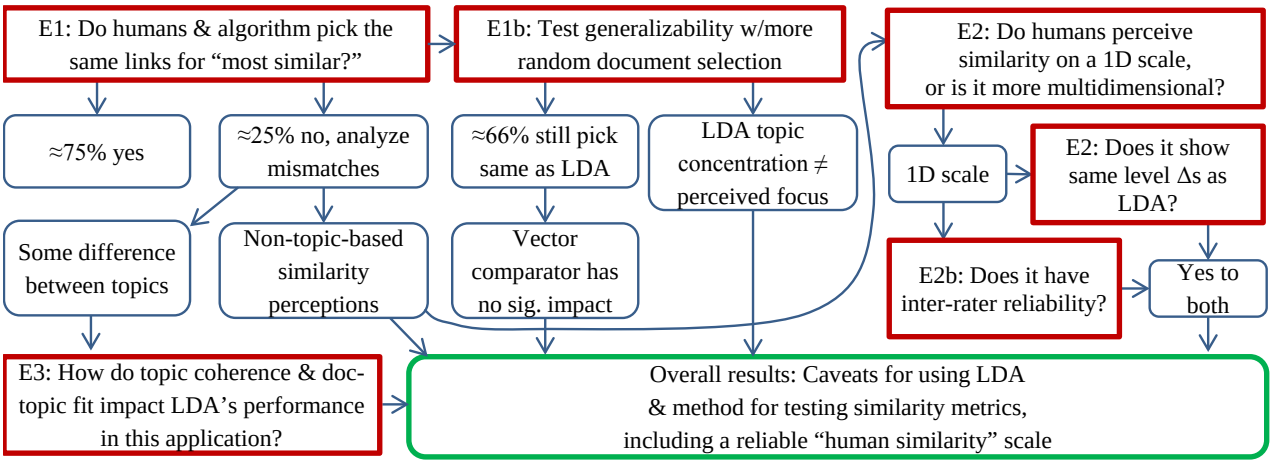

Figure 1: Overview of Experiments in Study 2

Qualitatively exploring the dimensions of similarity that people used, results from this study suggest that some ratings are not based on topical content. I explore these other bases of similarity judgments and the impact they have on human-algorithm agreement.

Based in part on these observations, I sought to explore human perceptions of similarity from multiple perspectives, producing a highly reliable multi-item human similarity measurement scale, further validating inter-rater reliability in a smaller test ("Experiment 2b."). I compared human similarity ratings of pairs of documents with algorithmic similarity scores and find generally high correlations.

In the first experiment within Study 2, when the humans and algorithm disagreed about choosing which two of three documents were most similar, those disagreements were not uniformly common across all of the topics. I sought to better understand what conditions or aspects of those topics might affect the agreement between algorithmic and human similarity. I conducted Experiment 3 to gain this understanding, and find that agreement is impacted both by topic coherence and the strength of fit between a document and its most closely associated model topic.

2.5: Study 2: Models And Materials Selection

2.5.1: Data

The input data set for this study comprises all 10,331 ideas submitted to the 2012 President's SAVE (Securing Americans' Value and Efficiency) Award, which at the time of study were publicly visible at saveaward2012.ideascale.com. U.S. President Obama began the annual SAVE award in 2009 as "a process through which every government worker can submit their ideas for how their agency can save money and perform better" (IdeaScale, 2012). Most ideas in the corpus are a paragraph or a few paragraphs, written by government employees during a short contest-based solicitation period. They span a wide range of quality and feasibility. Contributors have the opportunity to have their idea heard, recognized, and possibly implemented. Contest administrators read through the submitted ideas, selecting four finalists put to popular vote at WhiteHouse.gov/Save-Award, and the winner is brought to Washington, DC for presidential recognition.

I chose this data set because it is a real-world instance of the motivating example of ideation tools, which realistically represents the scale and diversity of results that might be expected from an ideation

process in a population the size of the US federal government. The motivation behind the ideation and award (reducing expenditures of public money) are generally noncontroversial, especially compared to other government-related topics, even while particular submitted ideas might be less neutral, as might be expected among contributions to other large-scale collaboration platforms. The corpus scale is comparable to the Stanford Dissertation Browser's data set of just over 9,000 thesis abstracts (Chuang et al., 2012).

2.5.2: Preprocessing

Documents used in this study each contain the title and body of one of the 10,331 ideas, excluding any comments, author information, or tags that were manually added in the original data. Knowing that the results of this study do not rely on such additional features, which may not be present in all other systems that can build on this work, increases generalizability. I set text to lowercase and removed stopwords before algorithmic processing.

2.5.3: Topic Modeling Algorithm

I chose a statistical technique that models a whole corpus of text, including its commonalities and themes. Such a technique supports comparisons between documents within the context of that corpus, which is appropriate for the task of generating within-corpus navigational links based on topical similarity. This is contrasted with representations used to compare documents independently of context, such as term frequency vectors. Probabilistic topic modeling such as LDA (Blei et al., 2003; Steyvers & Griffiths, 2007) is one such state-of-the-art technique, and widely used including in applications described above. More advanced techniques, including those not yet invented, might be evaluated using the human-centered methods described in this study.

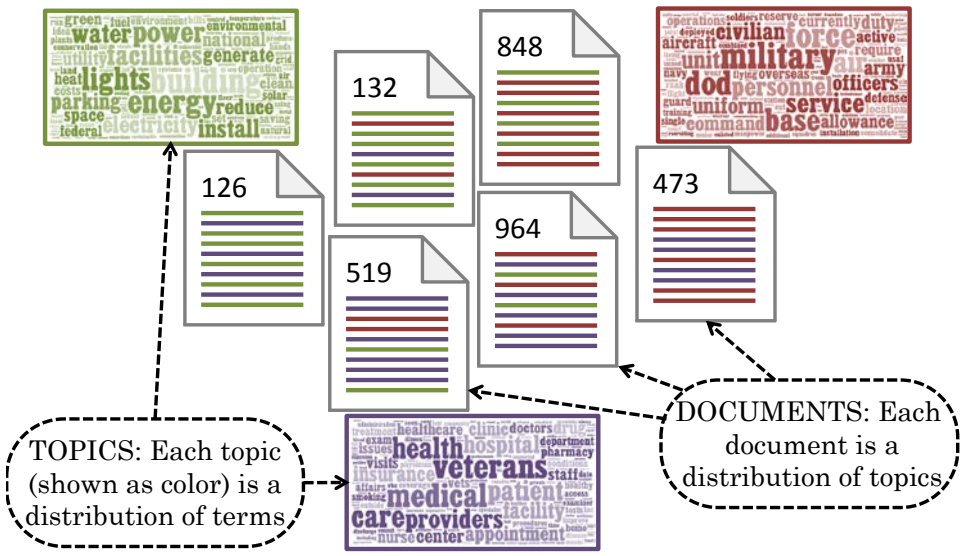

Figure 2: LDA generative model.
Each topic (here, a color) is a distribution of terms; each document is a distribution of topics.
Information embedded in the ordering of terms, sentences, etc. is generally discarded;
the model instead focuses on information about relative proportions ("bag-of-words").

As graphically illustrated in Figure 2, LDA models each "document" (shown as a page) as a multinomial distribution over "topics" (shown as colors) and each "topic" (shown in a colored box) as a multinomial distribution over terms. It assumes a model of document generation where for each term, a topic is first sampled from the topic distribution for that document and then a term is sampled from the term distribution for that topic. Users of LDA often assume that the distribution of topics provides a useful view either instead of or in addition to the word distribution. Similarity between two documents can be measured in terms of those topics (Steyvers & Griffiths, 2007), and this measure may be more accurate for documents that focus on those topics than for documents covering unusual content (Dinakar et al., 2012).

I ran an LDA model, using program default hyperparameters. These were $\beta=0.01$, as suggested in (Steyvers & Griffiths, 2007), and $\alpha=1$ which is in the same range as suggested by Steyvers & Griffiths (2007) and which does not bias the model toward smoothness nor sparsity (Steyvers & Griffiths, 2007).

As noted by Singh et al., "Selecting the right number of topics is an important problem in topic modeling" (2012, p. 143). Consistent with Xu and Ma's goal of maximizing dissimilarity between clusters (2006, p. 303), I selected the number of topics by choosing the model with the lowest average percentage of documents that has neither or both topics in each possible topic pair. This evaluation metric readily distinguished a 21-topic model from the tested range of 5 to 300 topics, as the model that maximally separated documents into different topics.

2.5.4: Sampling Strategy

My strategy for sampling documents was designed to serve several needs. First, the motivating application – linking documents in a corpus – suggests a focus primarily on how well high as compared to low similarity ratings match human judgment, because automatically generated navigational links are likely to be selected from the highest similarity ratings. I selected pairs from qualitatively different levels of "high" and "low" algorithmic similarity, as well as "low similarity with the same most-probable topic," to have clear differences between levels of LDA-based similarity and thus clarify interpretation of differences in experimentally measured results. I relaxed this in Experiment 1b, which tests for generalizability.

Intentionally choosing pairs with widely different levels of similarity makes results of this study less sensitive to particulars of the topic model, corpus, and similarity measure's distribution in the middle of its range. These larger differences between tested similarity levels increase the chance that human similarity judgments would agree with the algorithm, producing a rough upper bound on that agreement and providing an interesting set of failure cases (i.e., more clearly identifying where algorithmic and human similarity judgments diverged substantially). I investigated these failure cases to improve understanding of LDA's fundamental limitations. Experiment 1b steps back from the rough upper bound to provide a more general test that is more sensitive to mid-range values.

Second, I wanted to sample enough documents within each topic to investigate interesting within-topic phenomena, while having sufficient breadth across the range of topics in the corpus. This was also a goal of the sampling strategy used by Lee et al. (2005), in which a total of 50 news articles were chosen to represent a handful of topical clusters "in an attempt to ensure a broader spread of human judgments of document similarity" (Pincombe, 2004, p. 11).

In pursuit of this goal, for experiments other than 1b I chose documents strongly associated with a small but diverse set of distinct topics, as described in "2.5.5: Focal Topic Selection" below. This aspect of the method helps ensure that results of testing against human similarity on a subsample covers the breadth of topics covered by the corpus. (Experiment 1b does this by random selection.) This aspect also helps clarify interpretations in Experiment 3; had I selected topics by other criteria (e.g. randomly, as in Experiment 1b) I might have wound up with two or more similar topics and not been able to get as much unique value out of testing each one. When the marginal benefits (potential knowledge gains) from exploring an additional topic are proportional to how unique that topic is compared to what has already been explored, this diversity-based sampling strategy maximizes expected knowledge gain in a limited experimental budget.

Choosing documents most strongly dominated by those topics allowed me to more easily look for relationships between characteristics of topics and the degree of agreement between algorithmic similarity and human judgment, my third sampling need. I acknowledge that this strategy limits generalization away from documents that may not have a clear dominant topic in the model (effects explored in Experiments 1b and 3), but feel that the gain in clarity for how judgments about documents reflect on their dominant topics was worth the tradeoff. In Experiment 1b, I removed consideration of dominant topics from the selection process to test generalizability.

Finally, I wanted multiple human judgments per document pair in the experiment so that (A) I could better perform reliability assessment of my similarity measurement scale, and (B) the free text fields were completed multiple times, allowing me to separate common from idiosyncratic content. Accepting costs associated with human subjects experiments, that meant I would have to select a relatively small minority of topics and documents. Again, I judged this a tradeoff worth making: while reducing the number of topics or documents sampled in the experiment may reduce generality, this concentration increases my ability to look at particular judgments in much more detail and more reliably. This was the same decision and reasoning as used by Lee et al. (2005), for the same reasons (Pincombe, 2004, p. 11). This approach is complemented by the generalizability test in Experiment 1b, which optimizes for covering more content rather than getting repeated measurements of the same document sets.

2.5.5: Focal Topic Selection

Because a human subjects experiment does not allow me to efficiently explore all 10,000+ ideas in depth, I needed to choose a smaller number of documents to include in the test set. I wanted to ensure topical diversity while also having enough documents within each specific topic to be able to investigate any differences between those topics. This aspect of the selection strategy allowed me to test for and observe between-topic differences in Experiments 1 and 2 which I could investigate further, in Experiment 3. This section describes the method used to select a diversity of distinct topics.

Each topic is a vector of weights over the unique terms in the corpus. I computed the cosine similarity measure between these term-based probability distributions, and for each topic, computed its average cosine similarity with the other topics. I chose the five topics lowest on this measure, most unique from the other topics on average. In order, these five topics were about veterans' health care (topic ID #4), the military (#14), public social services (for food, education, etc.; #1), reducing building energy use (#3), and Social Security benefits (#20), as labeled by manual inspection of the topics' term distributions.

2.5.6: Focal Document Selection

Having selected topics as described in the previous section, I then selected "focal" documents from only the 27.4% of documents where the highest-weighted topic was one of those five. I focused on documents that were strongly dominated by the focal topics, as measured by the difference in probability between the highest-weighted topic (by definition, one of the focal topics) and the second-highest-weighted topic. For each of the five focal topics, I chose the five documents highest in this difference measure. This produced a set of 25 focal documents, each clearly associated with one of the topics picked in the prior step, so that the set of 25 focal documents represented a small set of distinct topics from across the data set.

Documents that were duplicates of other documents or had no understandable English content (one document) were eliminated and replaced with the next document on the list, iteratively. These quality filters caused 44 documents from one author, 11 from a second author, and 6 from a third author, to be disqualified from use in this experiment.

2.5.7: Computing Document-Document Similarity

2.5.7.1: LDA/Cosine

After selecting focal documents, I needed to select documents that the similarity algorithm computed were very similar as well as those the algorithm computed were not very similar, at qualitatively different levels, to be able to compare these differences with humans' judgment and measure whether or not the algorithmic measure matched human perceptions, at least at these coarse levels.

I represented each document in the corpus as a vector of 21 topic probabilities, which summed to 1 and was not smoothed (i.e. if a topic was not assigned to any tokens in that document, it had a probability = 0). I computed the **cosine similarity** between all possible pairs of document vectors. For

each document, this produced a scored ordering of the other documents based on the degree to which they contained the same distribution of topics.

For most documents, a few others were fairly similar, followed by many documents with lower similarity scores. Giving all pairs equal chance of selection would produce mostly low similarity ratings because of this "long tail." In order to follow my criteria of creating a task with clear differences in the LDA similarity measure, and also focusing on high similarity ratings, I first chose the document that was computed as *most* similar to each of the focal documents, labeling these as "near" documents. If the selection was a focal document or met the disqualification criteria stated above, it was removed and replaced with the next in sequence, as before. A single "near" document could be selected multiple times.[7] Since the task focuses on similarity judgments of a document pair, I felt that allowing a document to appear in more than one pair was not a problem.

For each focal document, I then separated the corpus into two groups: one group of documents containing the same highest-weighted topic, and a larger group of all other documents, simulating applications that might cluster or link documents primarily because they share a highest-weighted topic or category. To identify documents with clearly different levels of LDA-based similarity, I filtered to just the 20% (plus any documents tied with those in that bottom 20%) of each group that was *least* similar to the focal document (by the same measure), and randomly selected one document from each group. I call these the "distant/same lead topic" and "distant/different lead topic" documents respectively. These choices allow me to examine how much the single highest-weighted topic helps predict document similarity, which might simulate applications that link documents based on shared membership in a single category or cluster. The full experimental set includes 90 documents, with a median length of 39

[7] Two "near" documents were most similar to three "focal" documents each; five "near" documents were most similar to two "focal" documents each, and nine "near" documents were each most similar to a single focal document.

tokens (excluding stopwords). Experiment 1b removes the 20% filter and uses randomly selected "distant" documents, and as described in section 2.7.1 below.

*2.5.7.2: TF*IDF/Cosine Similarity*

As a bit of background, Walter and Back (2013) use cosine of TF and TF*IDF vectors, after stemming and stopword removal, to compute similarity of short ideas (average length 25 words) submitted on a crowdsourcing/ideation platform, as a basis for "second level" K-means hierarchical clustering. They analyzed over 40,000 ideas in 112 contests and claim that text mining can be used to help automate the long expert-driven process of submission evaluation, by identifying the most unique term distributions (equated to high quality). That paper gives evidence that selections based on uniqueness of term distributions are somewhat predictive of contest winners selected by expert judges (maximum F1 score .639). The authors suggest future work exploring "more sophisticated clustering" than simple term frequency (TF) or TF*IDF measures.

In addition to LDA, I also computed document-document similarity using two *feature vectors* alternative to LDA topic weights, based on the conceptually simpler unigram TF*IDF measure, with and without Porter stemming (as two different ways of constructing feature vectors).[8] TF*IDF, which is widely used, represents each document as a vector the size of the corpus vocabulary, where the weight for each of those values is directly proportional to how frequently the term appears in the document (TF) and inversely proportional to how frequently a word appears in the corpus overall (IDF). The latter term means that words that are common across very many documents have less weight. Words that are common and meaningless enough to be considered stopwords (a, and, the, …) have weight set to 0 by removal, which accelerates processing. When using these alternative *feature vectors*, I used the cosine similarity measure, like Walter and Back (2013). Using a feature set size as large as the

[8] I used interactive RapidMiner for TF*IDF and source-code Lingpipe for LDA implementation.

vocabulary means that when people use different terms to refer to the same concept (e.g. “aid,” “assistance,” “help”), the TF*IDF measure does not draw this connection, and when people use the same term to refer to multiple concepts (e.g. “bank,” “book,” “close”), a false connection is detected. Stemming reduces dimensionality by collapsing different forms of the same word, for example detecting a connection between one document referring to “service” and another referring to “services.” While TF*IDF is parameterless and conceptually simpler, other studies (e.g. Boyack et al., 2011) have found topic modeling to have superior performance.

2.5.7.3: Other Comparison Measures

Several efforts to improve on LDA topic modeling have been made, but as of the time of this work no specific advance seems to have yet gained comparably widespread adoption in the research literature, and to the extent that they are still based on LDA, most would be likely to share fundamental limitations of LDA explored in this study. The experimental methods described in this study can be repeated in future work with more advanced models (e.g. those based on word embeddings) to test new developments, though I would expect some of the inherent limitations of LDA (e.g. similarity judgments not based on topics) to still be observed in more advanced models unless the advance specifically addresses one or more of those limitations.

I also used the LDA feature vectors with four *measures for comparing vectors* alternative to the common and quickly computable cosine similarity measure. Cosine was the original method for comparing vectors even before LDA, as stated in the introduction of L**S**A (Dumais et al., 1988, p. 282), and is used in later neural-network-based word embedding methods (e.g. Boytsov et al., 2016; Brokos, Malakasiotis, & Androutsopoulos, 2016). After cosine, I used the symmetrized Kullback-Leibler (KL) divergence (1/2 the *J* measure of eq. 2.2 in (Lin, 1991)) with additive smoothing (adding .000001 to all values and rescaling) to overcome the issue that these divergences are not defined when one vector has some zero values. Third, I used the related symmetrized Jensen-Shannon (JS) divergence (1/2 the *L*

measure of eq. 3.4 in (Lin, 1991)[9]). Fourth, I computed Euclidian distance. Steyvers and Griffiths suggest all four of these methods[10] for computing similarity between documents (2007, p. 443).

In a related setting helping users navigate through a large set of online communities, Spertus et al. (2005) empirically found that the cosine measure showed the best empirical results when compared against other measures. Informed by that prior work, this study focuses primarily on the measure previously found to be best. In evaluation below, I find that methods other than cosine still yield similar results for the comparisons evaluated here (see sections 2.7.2 and 2.9.3 below for details).

2.5.8: Participants & Filters

Participants were recruited to the experiment website via Amazon's Mechanical Turk (MTurk), a paid microtask crowdsourcing platform. Since the documents discussed concepts specific to the US federal government, participation was limited to those who were in the United States according to their Mechanical Turk profile and a GeoIP lookup confirming access from the US for at least one participation in a given experiment, and who had had at least 500 assignments approved by other requesters (to help with quality control). I filtered out data from surveys that were not submitted as complete, as well as any in Experiment 1 where no pair of documents was selected. I also removed from analysis those by participants who submitted two or more blank or copy/paste form fields (idea summaries, or the explanations of similarity and difference) in the same assignment. I manually reviewed the fastest

[9] This is noted for full disclosure of exactly which formula I computed, but because I use only rank information, neither the ½ factor nor the square root transformation suggested by (Endres & Schindelin, 2003) should have any impact on my work. The ½ is suggested by (a) the concept of averaging two asymmetric divergence measures to create a symmetric measure, (b) Definition 1/Equation 4 in (El-Yaniv, Fine, & Tishby, 1997), (c) the definition at the top of p. 1860 in (Endres & Schindelin, 2003), and (d) mathematical derivation from equations 4.1 and 5.1 in (Lin, 1991).

[10] The printed edition of (Steyvers & Griffiths, 2007, p. 443) uses a different definition of symmetrized KL-divergence than the one (a) used here, (b) found in other sources, and (c) found in the version posted on an author's web site. I believe this is just a print error. This source suggests dot product as well as cosine similarity; I use only the latter (dot product scaled by magnitude) to focus on comparing documents based on their mix of topics, reducing variance caused by length differences between short ideas.

submissions, but did not find any compelling reason to remove further submissions based on speed in Study 2 Experiments 1 and 2. Work was discarded from any participant who, on both pages of at least one of their Study 2 Experiment 3 responses, gave the same answer to all questions.

On arriving at Experiments 1 or 2 for the first time, participants were asked a short page of demographic and background questions, each a 7-point Likert item plus "I don't know" for the latter two:

- I am knowledgeable about the US federal government.
- English is my native language.
- The US federal government spends money efficiently.
- The US federal government looks for opportunities to save money.

A standard 6-level educational attainment item followed. Experimental questions, as described below, followed on separate pages (see Figure 3).

At the end of all experiments, participants had an optional free text box for anything else they wished to say. The most frequent comments were variants of "thank you" and "no comments." Participants received a code for payment as advertised, generally US$.40 in experiments 1 and 2 and US$.15 for experiment 3, which had a proportionally shorter task.[11] Participants in experiment 2b, which was designed to test inter-rater reliability, received 90% of their payment in the form of a bonus after having completed all 15 instances of the task that were made available, as advertised.

In Study 2, Likert items in all three experiments were coded for analysis from 1-7 where 1=strongly disagree and 7=strongly agree. In analyses where items are dichotomized, the "agree" options are

[11] Because of the additional questions and slightly longer documents resulting from random selection, we increased pay by $.10 for most participants in Experiment 1b, with a 10x participation limit in that experiment.

coded as “1,” the “disagree” options are coded as “0,” and there is no neutral. All experiments except 1b used the same 90 documents selected via the process described above.

2.6: Study 2: Experiment 1

2.6.1: Experiment 1: Methods

My first experiment in this study explores human-perceived dimensions of similarity and the degree to which the algorithmic measure captures that perception of similarity, for use in the application environment of links for navigability. In Experiment 1, participants were presented with a focal document, its associated “near” document, and one of the two associated “distant” documents, chosen randomly. These were presented in random order but labeled A, B, and C in presentation order, as schematically illustrated in Figure 3. For each, participants were asked to summarize “What is this about?” as an incentive to read the documents carefully (Kittur, Chi, & Suh, 2008), and respond to two six-point Likert-style items: “I understood this passage” and “I think this is a good idea” with options of {strongly, moderately, slightly} {agree, disagree}. They were then asked to indicate “which pair of documents seems most closely related and different from the third,” analogous to the link creation task that is of greatest interest here. This task is often called “triading” when used in usability testing (Bank & Cao, 2015). The number of times that participants select a particular pair as being the most similar, as a percentage of the times they could have picked that pair, is a primary dependent variable of interest, and I experimentally compare it to the LDA-based predictions to see how well LDA performs.

In two free-text fields immediately following this selection, participants were asked (1) what the two have in common as opposed to the third, and (2) what makes the third different from the other two. These questions and all three documents were visible on a single Web page. On one following page, participants answered a pilot version of the human similarity scale used in Experiment 2, using the two documents they had just selected as being most similar.

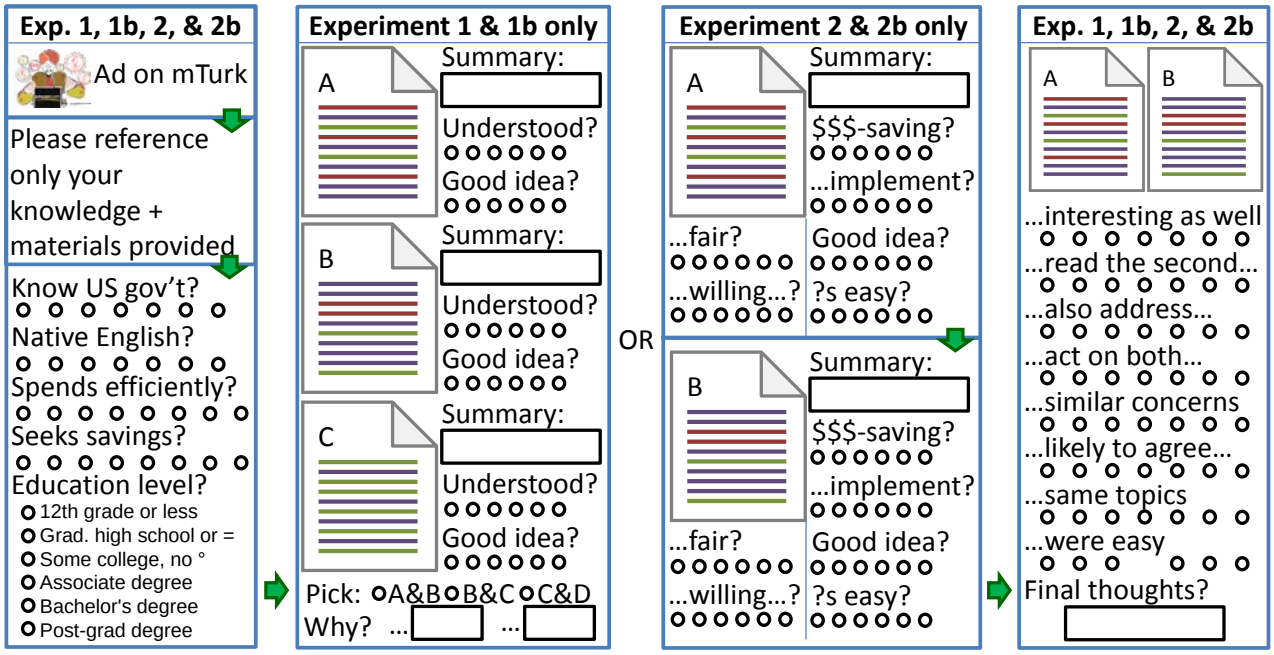

Figure 3: Schematic of primary pages showing page sequence & question summaries, Experiments 1-2b. Here, the last page shows that the participant selected A&B as most similar on the second page of Experiment 1.

2.6.2: Experiment 1: Topic-Based Similarity Determination?

To examine whether or not people used topic similarity as the basis for their judgments in this task, I examined the explanations people gave as to why the chosen two documents were similar and why the third one was different. A binary code was applied to responses, as a 1 if the primary distinction described that two were about the same topic and the third about another topic, e.g. "A&B talk about VA hospitals. [The other document] is about phones, which is an entirely different topic." If other reasons were given, such as understandability or idea quality, or where people focused on the beneficiaries of an idea, or its level of detail, scope, or the approach that the ideas were taking, that was coded as 0. For examples, "The first 2 are avoiding paying more money out. The third is a way to generate new income" shows a perception of similarity based on approach, and was coded as 0. "Brevity. The third is different from the other two in that it has a long and thorough explanation" shows a perception based on style, and was coded as 0. I coded all responses for analysis. To check for inter-

rater reliability, Dr. Rosé received codebook instructions and coded 115 response pairs, with 102 agreements and an acceptable Cohen's Kappa of .750. Results are shown in Table I.

2.6.3: Experiment 1: Main Results

In Experiment 1, 413 unique participants provided 562 results. Participants indicated that they used topics as their primary basis for perception of similarity in 72.8% of responses. As shown in Table I, just over 75% chose the same pair as the LDA-based similarity metric as being the most similar; a higher percentage agreed with LDA when choosing based on topics. This might be seen as an upper bound for this similarity measure's predictive validity in building links.

Those who did not choose the same pair as LDA were about evenly split between pairing the "distant" document with the "focal" or the "near" document. Interestingly, even among those who described the basis for their similarity choices as something other than topics, a slight majority still selected the same pair as LDA, suggesting that the topic model may still pick up vocabulary indicative of similarity that people do not semantically consider primarily about "topics" (e.g. beneficiaries or approaches). There were no significant differences in the demographic questions between people who chose the same pair as LDA and those who did not.

Table I. Responses to "Which of these two seem most closely related and different from the third?"
Percentages are of column totals. If responses were random, all cells would be ≈33%.

Document pair	Topic-based assessment	Other aspects	Overall
"Focal" w/"Near" (matches LDA)	344 (84.1%)	78 (51.0%)	422 (**75.1%**)
"Distant" w/"Near"	37 (9.0%)	35 (22.9%)	72 (**12.8%**)
"Distant" w/"Focal"	28 (6.8%)	40 (26.1%)	68 (**12.1%**)
Total	409	153	562

In the next two subsections, I analyze the possible roles of document understandability and perceived idea quality in explaining selections that did not match LDA predictions, also motivating other experiments as illustrated in Figure 1.

2.6.4: Experiment 1: Understandability

As illustrated by descriptions of why people selected a pair as most similar, some people chose the selected pair as similar because "They are easy to understand" while the other was different because "I have no idea what they're saying" or inversely, that the selected pair were similar because "they're both incomprehensible" and the third was different because "I actually understood what the person was suggesting." A rough coding indicated that about 25 of the description pairs were related to understandability, like these examples. Analyzing responses to "I understood this passage" revealed that most people understood most documents: of the 1666 responses to this question, only 3.3% were "strongly disagree" with less than 5% in each of the other "disagree" categories; the highest levels of understanding were the most frequent. Across the 90 documents, average understanding scores had a mean of 5.888 and a standard deviation of .94582. Free-text summaries indicated that even documents lower on this measure were reasonably well understood.

There were 69 cases where participants understood exactly one or two of the documents (as measured by the dichotomized "understanding" item), and the two matched their selection of which two documents were the most similar. More of these were cases where the person disagreed with LDA than would be expected if the variables were independent, by a Pearson chi square test (p=.016).

2.6.5: Experiment 1: Idea Quality

Other descriptions explaining participants' similarity choices focused on the idea quality or some aspect of it. For example, one response noted "A and B seemed like more thought out ideas" and the third is different because "It is not written well and is not a clear idea;" another wrote the inverse that the selected two are similar because "Both are harbrained crazy ideas that really don't make much sense" [sic]. Approximately 43 comment pairs appeared to focus primarily on idea quality. I then analyzed dichotomized responses to "This is a good idea" and found 119 cases where exactly one or two of the ideas were seen as good and the two matched their selection for "most similar." As with

understanding, a disproportionately large share of these were in cases where the selection disagreed with LDA, according to a Pearson chi square test (p=.045).

Of the 140 "most similar" pair selections that differed from the topic-based LDA prediction, 51 could potentially be explained by these measures of similarity in understandability or idea quality. This analysis helps us understand some factors that limit how well LDA can agree with human perceptions of similarity.

2.7: Study 2: Experiment 1b

2.7.1: Experiment 1b: Methods

To measure human and algorithmic agreement across a broader range, I also randomly selected 10% of the documents as focal documents for a generalizability experiment ("1b"). I did this by randomly shuffling documents and taking the first 1,033, skipping over documents disqualified as described in section 2.5.6 above. For the "near" document, I selected the most closely related document according to LDA, randomly choosing one of the four measures for comparing vectors discussed at the end of section 2.5.7 above. In this experiment, I randomly selected the "distant" document from among all nonfocal documents instead of from the most-distant 20%, so that the comparison between human and algorithmic similarity might be made over the range more broadly, and to remove any dependence on the algorithmic similarity measure for this part of the triad selection. I then repeated the experiment described above.

In this study's other experiments, I only selected documents that are strongly associated with a particular topic, for reasons described in section 2.5.4 above (similar to Lee et al. (Pincombe, 2004)) but which produces a rough upper bound on agreement. Randomly selecting focal documents helps assess agreement more generally by including documents that have a more "flat" or uniform distribution of topics. If this "flatness" means the documents are less focused, it may be hard for humans to compare with others. If, on the other hand, it means that the document is focused but the focus is not aligned

with the topic model, humans might be able to make reasonable comparisons even where the algorithm might have difficulty; it is at least a different range of values than was tested above. To help distinguish, I asked people to indicate their level of agreement/disagreement with "This focus of this idea is clearly expressed" as with understandability and idea quality.

To test whether or not participants' selections (of which pair in three is most closely related) depends on the method used to compare topic vectors and select the "near" document, I used Pearson's chi-square to test a null hypothesis that those two factors are independent. The chi-square test differs from other tests in that it can be used to confirm that a null hypothesis is correct (Privitera, 2015, p. 298).

Conventionally, assessments of Type II error (incorrectly retaining a null hypothesis) target an 80% power (probability of detecting an effect if present) (Cohen, 1992), considering mistaken rejection of a null hypothesis to be four times as serious as mistaken acceptance (Cohen, 1988, p. 5). Here, I target 95% power ($\alpha=\beta=.05$) and report the smallest effect size that I am 95% confident I would have observed if it existed, given post-filter sample size. That effect size, for the chi-square test, is reported in terms of an index *w* (Cohen, 1988, Chapter 7.2). The values of *w* corresponding to adjectives like "small," "medium," and "large" depend on the particular problem or field. For the psychology and behavioral science fields he was writing about, Cohen suggested that $w=.1$ corresponds to a "small" effect, $w=.3$ to "medium," and $w=.5$ to "large," cautioning investigators that these should be treated as a general frame of reference and not taken too literally (Cohen, 1988, pp. 224–225). Researchers often aim to detect a medium effect size of $w=.3$ (Newton & Rudestam, 1999, p. 76), which was intended to represent "an effect likely to be visible to the naked eye of a careful observer" and which has since been found to approximate the average size of observed effects in various fields (Cohen, 1992, p. 156).

2.7.2: Experiment 1b: Results

After filtering as described in section 2.5.8 above, I had 936 tasks completed by 458 Turkers, including "focus clear" assessments for 2752 documents. As shown in Table II, humans chose the same pair of documents as the algorithm **65.6**% of the time, with the remainder about evenly split between the other two options.

The hypothesis that participants' selections are independent of which method was used to compare topic vectors was **retained** with p=.320. A power analysis using G*Power 3.1.9.2 (Faul, Erdfelder, Buchner, & Lang, 2009) showed that this test would have a 95% chance of finding even a fairly small effect of size $w \geq .149$. The null hypothesis was also retained when combing the bottom two rows and with 95% probability would have detected an effect size with index $w \geq .135$. Based on these observations, I conclude that the method used for comparing topic vectors does not make a significant difference to which pair of documents participants selected as most closely related. I therefore combine all four methods for further analyses.

Table II: Experiment 1b results. Percentages are of column total.

Document pair	Cosine	KL	JS	Euclidian	Overall
"Focal" w/"Near" (matches LDA)	151 (62.1%)	145 (62.5%)	159 (71.0%)	159 (67.1%)	614 (**65.6%**)
"Focal" w/"Distant"	39 (16.0%)	43 (18.5%)	33 (14.7%)	39 (16.5%)	154 (**16.5%**)
"Distant" w/"Near"	53 (21.8%)	44 (19.0%)	32 (14.3%)	39 (16.5%)	168 (**17.9%**)
Total	243	232	224	237	936

After people chose which pair they thought was most similar, they rated the similarity of the pair along the scale described in section 2.8.1 below. For the document pairs that were rated, the human similarity score Spearman-correlated $\rho=.267$ ($p<.0005$) with the LDA cosine similarity of the two documents. As expected from differences in document selection, this is lower than the result reported in Experiment 2 below.

I compared average responses to "This focus of this idea is clearly expressed" with the "flatness" of the distribution of topics and found no correlation between what ideas people rated as less focused and "flatter" distributions. Here, "flatness" is measured by the standard deviation and/or Gini coefficient of the percentage weights in each document's topic vector. In both measures, numbers closer to 0 represent "flatter" distributions. As shown in the "word count" row of Table III, I also tested to see if longer documents (measured by total number of words assigned to any topic) were "flatter" or less focused. The absence of significant correlations in the first row of Table III should caution readers against interpreting flatter LDA topic distributions to indicate that a human would perceive the document to lack clear topical focus; a perceived focus might just not align with one of the relatively few "topics" dimensions the model has computed as primary for dimensionality reduction.

Table III: Spearman Correlation (ρ) Between Various "Flatness" Measures. *: p<.0005; n= 2752; n=10331.

	"Focus clear"	Gini	Standard deviation	Word count
Human: "Focus clear"	1	.004, p=.854	-.008, p=.674	-.011, p=.570
Gini coefficient	.004, p=.854	1; 1	.921*; .933*	-.460*; -.418*
Standard deviation	-.008, p=.674	.921*; .933*	1; 1	-.300*; -.272*
Word count	-.011, p=.570	-.460*; -.418*	-.300*; -.272*	1; 1

2.8: Study 2: Experiment 2

Experiment 2 uses a multi-item scale to measure perceived similarity, in order to test the reliability and distinctiveness of aspects of human similarity judgments.

2.8.1: Experiment 2: Methods

In Experiment 2, I randomly selected one focal document, and then with equal probability the "near," "distant/same lead topic," or "distant/different lead topic" documents, presenting the two documents across the tops of separate pages, in random order. Participants summarized each document, as before, and {strongly, moderately, or slightly} {agreed, disagreed} with each of the following statements:

- This idea is likely to save money.

- This idea would be easy to implement.
- This idea is fair.
- The relevant government decision-makers would likely be willing to do this.
- I think this a good idea.
- The questions on this page were easy.

The first four of these were intended to align with four commonly-used policy analysis criteria: effectiveness, efficiency, equity, and political feasibility, respectively. These were added after it became apparent (reading the free-text explanations of similarity) that some participants in Experiment 1 were using these factors to measure similarity between documents. The fifth question is conceptually similar to the four more detailed criteria (overall "good idea") and matched Experiment 1. Experiment 1's question about understanding was omitted, since participants were not choosing a pair of documents, so that pair selection could not be influenced by differential understanding.

Participants were then shown the two documents together, in the same order, and {strongly, moderately, or slightly} {agreed, disagreed} or were neutral with respect to the following statements:

- If I had just read the first idea and found it interesting, I would find the second idea interesting as well.
- It would be good for the author of the first passage to read the second passage.
- Addressing the ideas expressed in the first passage might also address the ideas in the second passage.
- It would be make more sense to act on both ideas together, than to act on them separately.
- These two passages address similar concerns.
- The author of the first passage is likely to agree with the second passage.
- These two ideas are about the same topics.

- The questions on this page were easy. [No neutral option.]

2.8.2: Experiment 2: Results

Questions were first analyzed individually. Each of the similarity questions is significantly correlated with the LDA-based similarity score for that pair (Pearson, $p<.05$). Using ANOVA as well as Welch and Brown-Forsythe statistics (which do not require homogeneity of variance) I found that each individual question has a statistically significant ($p<.05$) difference between ratings of pairs where both documents had the same lead topic as compared to pairs where documents did not share the same lead topic.

I then examined whether these questions addressed different dimensions of similarity or loaded onto a single scale. An exploratory principal components analysis showed that the seven questions asking in various ways about document similarity loaded on a single factor, with all items intercorrelated significantly ($p<.0005$) and strongly (average strength by Pearson=.578; by Spearman=.579). I analyzed those seven items as part of a single similarity construct, as suggested by Spector (1992), and found a Cronbach's α of .905 over the 457 surveys with answers to all these questions. This became my summated human similarity construct for further analyses (range: 7-49). For analysis of document pairs, I average the human similarity scores received for that pair.

In Experiment 2, the LDA-based distance between a pair of documents was controlled to three levels that are significantly different from each other. If this distance metric was measuring the same concepts as my scale, those same significant differences would be expected in the ratings. The human similarity measure was, indeed, significantly different across those three levels of LDA-based similarity. Further, the mean of the per-pair human similarity measure correlated significantly ($p<.0005$; $n=75$) with the LDA-based cosine similarity score ($r=.544$; $\rho=.553$).

In order to see if some topics produced higher human similarity scores than others, I tested to see if the per-pair rating was different across different focal topics. A difference was observed: Focal topics 1

and 20 led to significantly lower human similarity scores than the other three (p<.0005; 23.3 vs. 29.6 points; n=75). As a point of reference, the LDA-based cosine similarity score was not significantly different across document pairs in different focal topic groups. I investigate differences among the topics further in Experiment 3 below. From Experiment 2, I contribute a "human" measure of similarity between two documents, on a 7-item scale which has desirable properties.

2.9: Study 2: Experiment 2b

2.9.1: Experiment 2b: Methods

Experiment 2b was designed to measure inter-rater reliability for the scale in Experiment 2, because that experiment did not have enough examples where multiple people rated the *same set of pairs*, as most participants in Experiments 1 and 2 rated only a single pair and inter-rater reliability requires having different people rate the same set of materials. I made Experiment 2b an independent test by allowing only participants who had not been part of Experiment 2. The procedure was the same but used only 15 document pairs to be rated by all participants, randomly selected with balance across the five focal topics as well as the three LDA-computed similarity categories.

2.9.2: Experiment 2b: Results

31 participants each rated all 15 document pairs, for 465 additional results. Cronbach's Alpha for the seven-item scale in this data set is exceptionally high at .953 and would be lower if any item on the seven-item scale were omitted, giving me further confidence in inter-item reliability. The average of (31 choose 2 = 465) Pearson correlations (n=15) between raters' composite similarity scores is r=.747. Omitting the two raters whose scores sometimes did not correlate significantly with other raters, the average correlation is r=.790. All remaining correlations were significantly positive (p<.05; more than half with p<.0005). Therefore, I have confidence in the inter-rater reliability of this seven-item human similarity scale, even in the presence of noise from some raters.

2.9.3: Experiments 1 and 2: Joint Results

In Experiment 1, participants saw three documents and selected which pair was the most similar. They could choose the focal document paired with the "near" document, or the focal document paired with the "distant" document (which may have had the same highest-weighted topic or not), or they could choose the "near" and "distant" document pair. I computed the percentage of times each pair was selected, as a fraction of the times it was available for selection, as a dependent variable in this analysis. This is labeled "Human Pick 2/3" in Table IV.

In Experiment 2, humans rated the similarity of the focal document paired with either the "near" document, the "distant/same lead topic" document, or the "distant/different lead topic" document, and the mean score was computed for each of the 75 document pairs. This is labeled "Human Scale" in Table IV.

Jointly analyzing these two shows how the percentage of time that a pair was picked correlates significantly ($p<.0005$; $n=75$) with the human similarity metric ($r=.652$, $\rho=.623$) and the LDA similarity metric ($r=.772$, $\rho=.730$). As mentioned above, this can perhaps be interpreted as a rough upper bound on the degree of agreement between the LDA measure and human similarity judgments. This is also consistent with the Spearman correlation between human and algorithmic judgments of similarity that have been found at the word level (Faruqui et al., 2015). Table IV shows the correlation between these alternative measures and the human similarity scores for tested pairs of documents (as well as each other). Because the triads tested in the human similarity measures were constructed based on the LDA cosine similarity measure, those numbers are in **bold**.

I suspect that my basis for triad construction is why the "percentage of time that a pair was picked" measure from Experiment 1 correlates slightly stronger with the LDA measure than even with the highly reliable, independently measured human similarity scale of Experiment 2, and caution against drawing conclusions from this numerical difference. In the size of this dataset, and other data sets I hope the

learnings from this work can apply to, it is infeasible to collect human similarity judgments for all possible pairings as an input to triad selection. Lower correlations between TF*IDF and human perceptions of similarity, shown in *italics*, are consistent with prior work that finds LDA to have superior performance (Dinakar et al., 2012), but the correlations between human similarity scores and TF*IDF scores may have differed if I had selected the triads based on TF*IDF.

Table IV: Spearman Correlation (ρ) Between Various Similarity Measures. p<.0005; n=75.

	Human		LDA				TF*IDF, Cosine	
	Pick 2/3	Scale	Cosine	KL	JS	Euclidian	Unstemmed	Stemmed
Human Pick 2/3	1	.623	**.730**	.752	.745	.744	*.504*	*.577*
Human Scale	.623	1	**.553**	.547	.535	.480	*.441*	*.504*
LDA Cosine	**.730**	**.553**	1	**.980**	**.983**	**.919**	**.511**	**.639**
LDA KL	.752	.547	**.980**	1	.994	.936	.543	.665
LDA JS	.745	.535	**.983**	.994	1	.937	.520	.649
LDA Euclidian	.744	.480	**.919**	.936	.937	1	.525	.632
TF*IDF	*.504*	*.441*	**.511**	.543	.520	.525	1	.917
TF*IDF stemmed	*.577*	*.504*	**.639**	.665	.649	.632	.917	1

I then measured the extent to which my triad selection based on cosine similarity (see section 2.5.7.1 above) still tests coarse differences in similarity scores under alternative measures of comparing LDA topic vectors, and find that my selection of document sets reflecting coarse-level differences in the cosine measure also reflects comparable coarse-level differences in scores produced by these other methods for comparing vectors. The rank-ordering of similarity scores was not highly sensitive to the method used to compare vectors, as seen by the underlined correlations (average .958). For all four of the measures (Cosine, KL, JS, and Euclidian), the "near" documents were within the 0.62% of documents most similar to the focal document. Separating the corpus into one group of documents containing the same highest-weighted topic, and a larger group of all other documents, most of the "distant" documents fall into the half of those sets that was "least similar" according to each measure. There are only four exceptions: two of the "distant/same lead topic" documents under the KL divergence measure, and two of the "distant/different lead topic" documents under the Euclidian distance measure. These

observations about coarse level differences and rank-order correlation further support the conclusion from Experiment 1b that using another method for comparing LDA feature vectors does not make a significant difference in these results.

For the TF*IDF based measures, to which the separation in selecting "distant" based on "same lead topic" or not does not apply, I examined the positions of the selected "near" and "distant" documents in a list ordered by similarity to the respective focal document. The "near" documents used in this study's experiments had an average percentile ranking (where 1 is very similar and 99 is very dissimilar) of 20 (unstemmed) and 14 (stemmed). The "distant/same lead topic" documents had an average percentile ranking of 33 (unstemmed) and 32 (stemmed). The "distant/different lead topic" documents had an average percentile ranking of 55 (unstemmed) and 62 (stemmed). The differences between those levels is significant ($p<.05$) for the stemmed measure and the latter difference is significant for the unstemmed measure.

2.10: Study 2: Experiment 3

Experiment 3 evaluates the cohesiveness of topics and the perceived fit between documents and topics, motivated by observed differences between documents by most-probable topic above (see Figure 1). It then investigates the extent to which the performance of the topic modeling algorithm in this application context depends on the perceived coherence of the topics and the degree to which documents match the latent topics.

In Experiment 1, I found that which pair people chose as the most similar was not independent of the focal topic. I also found that of the 48 responses by 47 workers where documents in the pair selected as "most similar" did not have the same highest-weight topics, nearly half (23/48) came from three focal documents, and 44% (21/48) came from documents with two particular highest-weight topics (1 and 3). Combined with some of the free-text comments in Experiment 1, I was led to explore whether or not some of the cases when LDA did not predict the "pick two of three" task as well might be

due to less coherent topics or documents that did not fit well with those topics. As algorithmic metrics do not capture topic coherence well (Chang et al., 2009), I decided to use a human scale measure, and designed a direct experiment drawing from prior work (Newman, Lau, Grieser, & Baldwin, 2010). This experiment also tests Dinakar et al.'s (2012) observation that LDA-based algorithmic links seemed less helpful when they involved stories with unusual vocabulary which may not have clearly fit a single topic well.

2.10.1: Experiment 3: Methods

In this experiment, each of the five focal topics were shown in two different ways, assigned independently of topic: a word cloud with form similar to those in Figure 2, or a two-column list of the top 15 words and their probabilities, which included all words with probabilities >1%. Two visualizations were used in order to investigate the possible role of presentation in human judgments of coherence. Because results for both were virtually identical, I report only the aggregated results.

Participants were asked to come up with a short title for the collection, and then answered five items about topic coherence, each on a six point agree/disagree scale:

- Coming up with the title was easy.
- This collection of words is coherent.
- If I put these words into a search engine, it's pretty clear what the returned documents would be about.
- These words all belong to the same topic.
- This collection of words is meaningful.

Participants could complete the task up to 5 times, seeing a different topic each time but always the same representation, randomly assigned on their first visit.

On a second page, they were shown the same set of words in the same form, shown one of the short documents, and asked to answer another six items on the same six-point agree/disagree scale, the first five measuring how well the document fit into the topic:

- If I searched for these words, I would expect to find this document.
- If someone subscribed to the collection or feed of documents that generated this list of words, it would be appropriate to send them this document.
- This document fits into the topic described by the list of words.
- I would expect to find these words in other documents related to this one.
- This document belongs in the same collection as the documents that generated this word list.
- The questions on this page were easy.

The document shown was either one of the focal documents associated with the displayed topic, or one of the "near" or "distant" documents that had been presented with that focal document in Experiment 1. I expected (A) that the "distant /different topic" documents would not be seen as fitting in well with the focal document's highest-weighted topic, (B) that selected documents might on average be seen as fitting less strongly with less cohesive topics, and (C) that if some of the documents did not fit particularly well with their focal topics, this might affect how people selected the "most similar" pairs in Study 2 Experiment 1.

This experiment draws on Newman et al.'s (2010) human evaluation of topic coherence for comparison to automated methods of evaluating topic coherence, but uses a much larger crowd of raters instead of trained specialists. Participants evaluated coherence using the five-item scale above, drawn from Newman et al.'s description of their measure.

2.10.2: Experiment 3: Results

After the filtering described above, 309 unique workers provided 993 complete results, for an average of 11 ratings per document in Study 2 Experiment 3. A reliability analysis similar to that of Study 2 Experiment 2 was done on each of the two constructs in Study 2 Experiment 3. The "topic coherence" scale had a Cronbach's Alpha of .839 after removing the "coming up with the title was easy" item, which was not as strongly correlated with the other items as suggested by my source for the scale (Newman et al., 2010). The five-item topic-document fit scale had a Cronbach's Alpha of .976. That would have been slightly reduced by removing any item, so all were retained.

To explore document-topic fit, I dichotomized the mean document-topic fit score around its neutral point of 20. All but two of the focal documents (both in topic 1), all but one of the "most similar" documents (from topic 1), and all but eight of the "distant/same lead topic" documents (half of those eight from topic 1) were seen as fitting into the highest-weighted topic. This human measure of document-topic fit correlated $r=.506$ ($p<.0005$, $n=65$) with the fraction of a document's tokens assigned to its dominant focal topic. 24/25 of the "distant-different" documents were seen as not fitting into the focal document's highest-weighted topic. To the degree tested, the vast majority of documents fit their most-probable topics well, and did not fit other topics well, as expected. One topic accounted for most of the exceptions, with significantly lower document-topic fit ($p<.0005$, $n=672$, contrast value of -5.23, on a scale from 5 to 35), and it was one of the topics with significantly lower agreement with LDA and lower human similarity scores, above. Manual examination suggests that this topic covered a wider range of concepts (within Health & Human Services) than the other focal topics; it also scored significantly ($p<.0005$) lower than other topics on the coherence scale.

In general, there was a positive association between whether or not a participant's selection in Study 2 Experiment 1 matched LDA's, and how well the focal document was perceived to fit its highest-weighted topic. The mean document-topic fit score was 27.2 in cases where people agreed with LDA

and 24.6 when they did not, a statistically significant difference ($p<.0005$; $n=562$). The dichotomized document-topic fit for the focal document is not independent of whether a person selected the same pair as LDA ($p<.0005$).

Limiting analysis to cases where the "distant" document had the "same lead topic" as the "focal" and "near" documents, participants were more likely to disagree with LDA and more likely to choose the focal and "distant" pair when the "distant" document fit well into its topic. In a majority of the cases where the participant chose the "focal" and "distant/same lead topic," the "distant/same lead topic" document was seen as a better fit with that topic ($p<.0005$).

In 416 of the 562 cases (74%), Experiment 1 participants selected the two documents most closely associated with the focal topic. In 28 of the 140 cases where participants disagreed with LDA, their selection was consistent with choosing the two that best fit the focal topic (according to ratings from Experiment 3).

2.11: Study 2: Limitations

I have done this work on only one data set, and most of the work was done with a subset of that corpus constructed to compare pairings that have very high levels of LDA-based similarity scores with pairings that have lower algorithmic similarity scores, based on the motivating application. Except when testing generality in Experiment 1b, I deliberately chose conditions favorable for observing agreement between humans and LDA, such as selecting documents with a single very dominant topic and selecting a number of topics that maximally distinguished between documents. I did this in order to see how well LDA could perform under favorable circumstances – roughly, an upper bound – and to generate exceptions for further investigation about what LDA seems unable to handle.

2.12: Study 2: Conclusions

In Study 2, I have presented a method for evaluating a measure of document similarity, including a "choose two most similar of three" task setup as well as an independent human similarity rating on a novel and reliable seven-item scale that captures a single conceptual dimension of similarity. The methods for preparing data and testing it against an algorithm, described above, can be adapted for use with other textual data sets. The methods can also be used to evaluate more advanced topic models or other similarity measures.

I have applied these methods to evaluate an LDA-based cosine similarity measure. I found that for the most part, under deliberately-chosen conditions favorable for observing agreement, human similarity judgments and the "simple" LDA-based cosine similarity measurement often (≈75%) agree, and that agreement is closer to 66% when documents are selected more generally. There is still room for improvement of the algorithm in being able to automatically generate links that match selections humans would make.

Although I built these experiments using one specific method for comparing topic vectors (cosine similarity), the rank orderings produced by other comparison methods are quite similar. More importantly, the limitations of the LDA model that I found and explored here are limitations of the LDA model regardless of the approach used to compare feature vectors, and regardless of certain improvements that might be made on the basic LDA model. Even a highly refined factorization or data reduction model attends to some features (e.g. topics) and obscures others, and it is important for practitioners who hope to help and guide human readers with an algorithmic similarity measure to understand those limits better than what prior work permits.

Exploring human perceptions of similarity more deeply, I qualitatively and quantitatively identified aspects of documents, such as understandability and idea quality, that LDA does not pick up on, nor could it be expected to, from its "bag of words" model. I found that at least in this setup, topic

coherence and perceived fit between documents and highest-weighted topics appear to be important factors in predicting what humans will choose on the link analogy task; such fit may be generally important when using dimensionality reduction techniques. These are caveats for the use of LDA for link generation based on topical similarity.

This work provides support for the value of working on techniques that can correct for the observed shortcomings, for example by incorporating features beyond word or topic vectors (such as tags, authors, connections between authors, etc.) in measuring similarity, using sources of prior knowledge to constrain the topic model (e.g. Andrzejewski & Zhu, 2009; Andrzejewski, Zhu, Craven, & Recht, 2011; Zhai, Liu, Xu, & Jia, 2011) or using a human-in-the-loop mixed initiative approach (e.g. Huang & Mitchell, 2007), so that the dynamic structuring of content benefits from both automation and human knowledge.

Future work exploring these topics further could explore how human measurements of topic coherence match algorithmic measures of topic quality, such as the "flatness" of the term distribution that makes up the topic, expanding on Chang et al. (2009). If an algorithm can automatically understand when its similarity measure will perform well vs. poorly, it can more appropriately add/not add links and structure, differentially weight ensemble methods, and better incorporate human feedback, leading to more effective use of these large-scale collaboration support tools.

Study 3: Evaluating the Effects of Exposing Details of Deliberative Process on Readers

3.1: Study 3: Chapter Summary

Large-scale collaboration systems often separate their content from the deliberation around how that content was produced. Surfacing this deliberation may engender trust in the content generation process if the deliberation process appears fair, well-reasoned, and thorough. Alternatively, it could encourage doubts about content quality, especially if the process appears messy or biased. Study 3 is a set of experiments attempting to distinguish between these competing possibilities.

Participants answered preliminary questions and then read a provided segment of a Wikipedia article on one of four randomly-assigned controversial topics with high-quality articles. I then showed participants a section of a discussion (vignette) related to the part of the article they had just read. The discussions were constructed to show various kinds of discussions that happen behind Wikipedia articles, including discussions with and without conflict, and with different conflict resolution strategies. Participants were then asked to rate the quality of the *article*. Between-subjects analysis showed that people who had read the *exact same* article rated its quality differently depending on what (if anything) they had seen of the related discussion.

This study found that surfacing deliberation generally led to decreases in perceptions of quality for the article under consideration, especially – but not only – if the discussion revealed conflict. The effect size depends on the type of editors' interactions. Finally, this decrease in actual article quality rating was accompanied by self-reported improved perceptions of the article and Wikipedia overall.

3.2: Study 3: Introduction

Technological advances in recent years, such as wikis, have enabled large-scale systems for aggregating knowledge and information from a very large number of participants, who bring a broad range of viewpoints, information, and expertise. In some systems, like Wikipedia, a significant amount

of coordination communication is employed to effectively combine these contributions (Kittur, Suh, Pendleton, & Chi, 2007).

However, many readers of Wikipedia are unaware of this work that has gone into the creation of an article, even though these efforts represent a potential source for readers to understand and evaluate the trustworthiness of an article. For example, imagine a reader who encounters a controversial topic in Wikipedia. In one case, the reader sees only the article and must evaluate the likely bias and validity of the topic on its own. Alternatively, the reader also sees that substantial and considered discussion has taken place among the editors of the topic on how to sensitively and appropriately present it. In the latter case, the reader has additional information to judge the article quality. If the discussion seems measured and fair, the reader's evaluation of the article may improve. If, on the other hand, the reader sees unchecked biases or personal attacks among the contributors, the reader may be less likely to trust the content than if it were encountered in isolation. Readers may simply become overwhelmed by the amount of information needed to understand the article creation process. More fundamentally, increasing the visibility of the uncertain and messy process by which articles are created may undermine readers' perceptions of trust — even if that process leads to a preferred outcome (Simon, 1996, pp. 1–14, 111–138). To quote John G. Saxe, "Laws, like sausages, cease to inspire respect in proportion as we know how they are made" (1869). The same may be true of user-generated content; the hypothesis that it is is hereby termed the "sausage hypothesis."

To examine this question, I conducted an experiment in which I surfaced various types of discussions along with content, and measured readers' perceptions of the quality of the article excerpt being discussed. I found that when discussion was provided alongside content, the quality ratings for the content were significantly lower than when no discussion was displayed, supporting the "sausage" hypothesis. When discussion involving conflict was displayed, article quality ratings were even lower. However, if the editors involved in the conflict resolved it through a positive collaboration approach, the

negative effects of conflict disappeared. Participants were not generally aware of the rating-lowering effect of the discussion, and generally reported that reading the discussion raised their perceptions of both the article's quality and Wikipedia in general.

3.3: Study 3: Conflict Resolution Strategies

Conflict has been studied from a variety of different perspectives, mostly in the context of small groups. Thomas (1992) and Kilmann described a framework in 1976 characterizing approaches to managing conflict, by the degree to which individuals attempt to satisfy their own concerns ("assertiveness") vs. others' concerns ("cooperativeness"). They find the five distinct conflict management strategies shown in Figure 4.

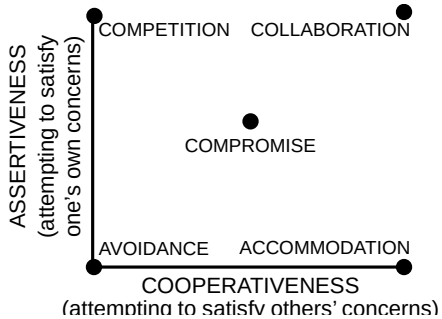

Figure 4: Thomas's two dimensional taxonomy of conflict handling modes

This framework has been adopted in the literature and validated several times (Thomas, 1992, p. 269). It is supported by a number of other studies that independently attempt to build typologies of conflict resolution strategies and identify aspects of conflict management that are important to outcomes, such as Klein & Lu (1989)'s early analysis of approaches to solving task conflicts in a cooperative interdisciplinary design team. Klein & Lu's primary conclusions are that "conflict resolution plays a central role in cooperative design…and knowledge acquisition in cooperative design presents special challenges and requires special techniques." The authors believe that new practices with parallel interaction amongst diverse concerns cannot happen without effective conflict resolution (Klein & Lu, 1989, p. 169).

In this study, I examine how revealing details of the joint production process affect perceived quality of outcomes. Task conflicts have a wide range of potential effects, both positive and negative. Therefore, I explored only task conflicts, as opposed to relationship conflicts which are generally only seen to have negative effects (see e.g. Jehn, 1997).

3.4: Study 3: Wikipedia Talk Pages

Wikipedia is an open and free encyclopedia which anyone can edit. The encyclopedic content and editing process are well studied as an example of a large-scale open collaborative system. A "discussion" tab in the upper left corner of most article pages links to a "Talk" page where editors discuss changes to the article (Wikipedia contributors, 2012). In 2013 and continuing through at least 2015, the Wikimedia Collaboration team was in active development on (and from mid-2015 through at least early 2017, actively maintaining) an alternative system called "Flow" for meaningful conversations in support of collaboration, which is both more accessible for new users and more efficient for experienced users (Bisson et al., 2015).

Viégas et al. (2007) confirm Kittur et al's finding (2007) that Talk and Project pages are the fastest growing parts of Wikipedia, especially with more heavily edited articles. They coded subsets of 25 article Talk pages (excluding archives) to find out what people are doing there, producing a typology which guided my choice of discussion types. They chose controversial and non-controversial topics in areas from hard science to pop culture, especially including cases with difficult coordination issues. Postings were coded for 11 binary dimensions, which were analyzed for frequency.

The most common kind of posting that work found is "requests for coordination," with contributors asking for help and explaining why they think specific changes should be made. Over half of Talk page contributions fit this category, including 97% of the discussions coded from the Yasser Arafat page.

The next most common use was a "request for information," found in just over 10% of posts. Writers of these posts hope to tap into the knowledge of an "approachable community of experts" (Viégas et al., 2007, p. 8) on a specific topic, without necessarily having intention to edit the article. The overwhelming majority of these requests were answered, with information or links that might answer the question.

The third most common type of posting was coded as "off-topic remarks," generally users sharing trivia or personal experiences related to the article topic. The fourth most common type of posting, with 7.9% of Talk page activity, includes references to official Wikipedia guidelines, as guidance for article editing. This pattern generally followed "serious disagreements or flame wars" in response to high levels of conflict (Viégas et al., 2007, p. 8).

3.5: Study 3: Judging the Credibility of Information Online

A number of studies have examined factors impacting reliability perceptions of online sources, including Wikipedia. For example, Fogg et al. (2003) studied over 2600 participants evaluating 100 real-world websites in 10 categories, identifying 18 top features that people consider when evaluating Web site credibility. Wikipedia fixes some of these to be the same across all articles, such as Design Look, Structure, Advertising, and Site Functionality. Other of Fogg's factors vary between articles, such as perceived information accuracy and bias, tone of writing, author motive, and readability; the discussion behind an article sometimes reveals signals about these factors. I hold most of these factors constant to focus specifically on how the presence and type of discussion influences perceptions.

Lucassen and Schraagen (2010) build on this work with an experiment to discover "which elements of Wikipedia articles university students use to assess their trustworthiness." They found the major elements to be textual features such as comprehensiveness, correctness, length, and pictures. I hold these factors constant and extend the investigation to another factor, examining how the discussion behind content impacts trustworthiness evaluations when it is used. Chesney (2006) further suggests that people will rate content they are more familiar with as more credible than content on a random topic, which I observe but control for by random assignment.

Stvilia et al. studies a more objective information quality measure, noting early on that "The same information can be judged as being of different quality depending on the context of a particular use" (2008, p. 983). This work describes a correlation between "actual" quality as measured by Featured

Article status and discussion pages that are large, readable, and well-organized, but that translation of interest to this quality measure depends on the content of the discussion and whether or not a consensus has been reached (Stvilia et al., 2008, p. 992). Differences in collaboration patterns are known to lead to differences in actual quality of an article as measured by "Featured" status (Kittur & Kraut, 2008; J. Liu & Ram, 2011). I would like to know if changing what somebody sees about the discussion correspondingly changes their perception of the article quality. Perceived quality is important for establishing legitimacy and building a more active community.

Stvilia et al. (2008) call for empirical studies of information quality, suggesting the English Wikipedia as a particularly interesting case for study. I answer that call and provide empirical evidence for some of the paper's observations, noted above. That paper partially defines information quality as "noncontroversial," assuming greater capacity to objectively evaluate articles. My extension to controversial topics serves as a building block for online deliberation systems that may eventually help people solve complex problems where objective evaluation may be impossible or less important than perceived quality.

The way that discussion around information is presented to users is an important design choice for an online community, with important consequences for how accurately the community is able to judge the credibility of information shared there. It may be, for example, that hoaxes are more likely to lie undiscovered when discussion is separated from content, as in Wikipedia, compared to sites where discussion is front and center, such as Reddit, and that hoaxes are more likely to propagate in systems where the discussion is decentralized (as in Facebook as compared to Wikipedia or Reddit) (Appelbaum, 2012; Garber, 2012). It is important to understand how the visibility of discussion may change perceptions of quality in the system, and how the effects of salient discussion may depend on the content of that discussion.

A number of experiments in adding visible discussion capabilities to sites that previously did not support this kind of interaction around specific content have been tried and shown to add value. For example, Kriplean et al. (2012) focus on grounded discussion and active listening, discussing the content even on sites that are already forums, as a means of increasing empathy and positive participation while efficiently summarizing and clarifying content.

Kittur, Suh, and Chi (2008) cite significant distrust in user-generated content in current large collaborative systems, and hypothesize that readers lack sufficient information to evaluate the trustworthiness of content. They present a "Wikipedia dashboard" with visualizations of churn, reversion, and editor registration, showing who edited how much when, and found that the information raised or lowered trust ratings, depending on whether the record implied an irresponsible (e.g., many anonymous edits) or responsible process. Shneiderman (2000) recommends full disclosure of sources' past performance patterns "in comprehensible and compact terms" as a means of building trust. Viégas, Wattenberg, and Dave (2004) showed a history flow visualization to reveal conflict and collaboration in article histories. Pirolli, Wollny, and Suh (2009) extended that work with a more comprehensive view of edit histories, displaying only real data for each article in their lab experiment. They also demonstrated significant increases in article credibility ratings through exposure of editor identity and more detailed histories.

I build on this line of work by explicitly examining how readers' opinions of quality and trustworthiness may differ based only on what is exposed about the discussion that led to the content, termed content transparency (Stuart, Dabbish, Kiesler, Kinnaird, & Kang, 2012). In particular, I addressed the following research questions:

3.5.1: RQ1: What is the effect of exposing discussions about article content on perceived article quality?

Since discussion about controversial topics can reveal disagreements, inaccuracies, and possible bias, or remind readers that fallible non-expert editors created the content, revealing discussions could significantly diminish perceptions of quality (the "sausage hypothesis"). However, seeing the discussion could also show that content is being discussed and vetted at some level, which might increase perceptions of quality.

3.5.2: RQ2: Do different kinds of conflict resolution have different effects on perceptions of content quality?

Some aspects of the discussion, such as inequality in editors' experience levels, are known to lead to differences in rated quality, while other aspects, such as inequality of contributions, do not (Arazy & Nov, 2010). I investigate how differences between discussion strategies also influences rated quality.

Resolving conflict by personal attacks, threats to leave the community, or ignoring complaints seem likely to have much more of a negative effect than resolution accomplished more rationally, for example by citing policies or sources. I expected that conflicts resolved through collaboration would be seen as more likely to increase the quality of the resulting article text than other approaches to conflict resolution.

Montoya-Weiss, Massey, and Song (2001) provide some hypotheses for how conflict resolution strategies in internationally distributed groups communicating asynchronously through text affect actual performance on a marketing consulting project with some time pressure. They found that avoidance and compromise behavior (as experienced by team members) significantly hurts performance, accommodation has no effect (because the text channel may not have been expressive enough to make this strategy obvious for team members to experience), and competition & collaboration correlate significantly and positively with performance. Montoya-Weiss et al. (2001) found the compromise approach may have hurt performance because of its manifestation, cutting and pasting possibly

contradictory content from different team members into a final team document, without integration effort.

3.5.3: RQ3: What do participants believe about how viewing the discussion may have changed their perceptions?

While RQ1 and RQ2 compare quality ratings from large groups of people who see different types of discussion or no discussion at all, RQ3 examines how individuals believe they are affected. In addition to their ratings of this particular article, I would like to see how the presence of discussion impacts participants' perceptions of Wikipedia overall.

This part of the work seeks to discover whether the effects of different types of discussions are brought about by a deliberative process (what Kahneman (2011) calls System 2) or by a more automatic associative process (Kahneman's System 1), and how these two mental systems interact to evaluate perceived quality. People are generally able to report with reasonable accuracy on System 2 activities, while associative activities generally escape awareness. A common experimental approach used to investigate especially System 1 psychology is to perform a between-subjects controlled experiment that demonstrates reliable differences caused by the manipulated variable, but where the cause of the effect never enters conscious thought (System 2), and participants may believe they were not affected or affected in the opposite direction. If the observed results of RQ3 and RQ1 or RQ2 do not align, at least part of the explanation must lie in what System 1 processes without the involvement of System 2.

It is possible to change the way people cognitively evaluate information, and Kahneman presents several factors known to more readily engage System 2. Simply presenting arguments has been shown to increase participants' cognitive evaluation of case descriptions in experiments evaluating the perceived legitimacy of Supreme Court decisions (Mondak, 1990). I extend that work by examining a source with a very different base level of credibility, examining how presentation of arguments (of different styles) may change cognitive evaluation and perceptions of quality.

3.6: Study 3: Experiments

3.6.1: Overview

I showed participants a segment of a collaboratively produced article and then showed them either no discussion or one of ten different discussions about the article segment. I then asked comprehension questions about the article and discussion to promote deep reading. I then asked participants to rate the article and the discussion on a number of scales, including ratings of article quality and the perceived level of conflict in the discussion. I collected data about demographics prior to the task, and collected self-report data about how the participants thought they may have been affected by seeing the discussion at the end.

3.6.2: Data Source

I chose Wikipedia as a ready source of data that has both a wide range of content along with publicly available discussion and resolution of disputes about the content. I selected topics that were both controversial and had high-quality articles. I wanted controversial topics with many disputes so I could identify a number of instances of different kinds of resolution. I also wanted articles that were high quality, because low quality articles are more likely to have features visible in the writing — poor style, lack of clarity, etc. — that could dominate judgments about quality, giving the manipulations less of a chance to have observable effects.

To identify controversial topics, I looked at community-curated lists of controversial issues and pages that were explicitly tagged as controversial, displaying and linking to an appropriate notice. I only considered articles in the main project namespace. This excludes, for example, Wikipedia policies and coordination pages, and uploaded files. This process identified 3403 unique controversial articles.

I then looked for examples of articles where the collectively generated content had reached a high level of quality. As with "controversial," I looked to the Wikipedia community's identification of the highest quality articles. The community uses an assessment scale to measure article quality, on an

ongoing basis through a manual process. I selected Featured Articles, which are "the best articles Wikipedia has to offer, as determined by Wikipedia's editors" (1869). This measure correlates significantly with quality assessed by external raters (Kittur & Kraut, 2008). At the time when I copied the list, there were 3189 featured articles (<0.1%). 50 articles were listed as both "controversial" and "featured," and I selected those for further investigation.

I examined this group of 50 articles and found common topical categories: Health, Science, Religion, Politics & History, Pop Culture, and Places. I chose articles across these categories, to ensure diversity in the pool of topics. I did my best to choose topics that were not currently dominating discussions in the news, and would be unlikely to change much over the course of the study period, because news reporting during the experiment could have unpredictable effects on the results. The articles selected were Autism, Pope Pius XII, Yasser Arafat, and the Cretaceous–Paleogene extinction event.

I read through the Talk page archives of these articles to determine the specific controversies present in that article. Some issues came up many times in different Talk page discussions, and the discussions were manually open-coded for the primary topic of discussion. Then, an issue was chosen based on its presence as an important controversy in that article, but which would be unlikely to be affected by participants' strongly charged and very diverse viewpoints. As an example, in the Yasser Arafat article, I chose the controversy about his place and date of birth (a common controversy for celebrities) rather than the ones around his sexual orientation or the use of the word "terrorist." I did this to reduce variance and maintain a relatively constant level of emotionality, since high emotionality in conflict has been shown to lead to lower quality in conflict resolution outcomes (Jehn, 1997).

For each article, I then created ten brief vignettes illustrating Talk page discussions about the particular chosen controversy, based in part on the styles (and some original content) from the discussions I observed. I portrayed a conflict between editors and resolution through each of the five

approaches described by Thomas (1992). Two pilots of this experiment included only conflict conditions, and found that revealing strong conflict among editors lowered perceptions of article quality. To investigate rating differences beyond those that may be caused by seeing conflict, I also composed four non-conflict discussion conditions, three of which were based on the three most common Talk page uses described by Viégas et al. (2007). The fourth shows one editor reporting on changes s/he made to the article, with no apparent controversy; a second editor leaves one word of thanks. Also based on Viégas et al. (2007), I had a sixth conflict condition depicting a single editor changing an article after removing a source that was inaccurately cited, implicitly addressing Wikipedia's policies around reliable sources; the editor is a frequent Wikipedian who also uses an abbreviation "POV" derived from the acronym for Wikipedia's "Neutral Point of View" core content policy (Wikipedia community consensus, 2012a). I classified this as a conflict condition based on Viégas's description that this strategy is used as a response to conflict (p. 8). A manipulation check later verified that this condition "behaved" like a conflict condition with respect to the measure of perceived conflict. These vignettes were originally created as a set for one article and then adapted for each of the other topics by substituting in the topics, sides, and sources used in discussion, aiming to maintain similarity of structure and conversational style.

Some of the discussions contained excerpts from actual interactions, though all the discussions presented were created by me to represent a certain discussion type. These discussion vignettes were written with a number of principles in mind. All were written to be short to minimize the time and attention required of participants. The vignettes were written to clearly demonstrate each of the discussion types, and were similar across article topics (within the same discussion type). Because quotes were not direct, editors' names were replaced by two-letter initials, and each topic showed discussions among nominally different editors. Links and other aspects of Wikipedia formatting were retained in the presentation of both the article and discussion segments. Links to references were

retained, but the actual references were not shown, to focus participants' evaluations on the text itself. No pictures were included in the short segments of this experiment. The actual discussions used are available in appendix A to (Towne et al., 2013).

3.6.3: Experimental Design: Details

As suggested by Kittur et al. (2008), I posted a request on Amazon's Mechanical Turk requesting participation in this experiment and describing the task. Turkers who accepted the task were randomly assigned to an article topic and discussion type, neither of which they had seen before in this study. I restricted eligibility to Turkers who had at least 95% of their prior tasks approved, and who were in the United States (according to their Turk profile).

In all conditions, participants were first asked some preliminary questions about their background and use of Wikipedia. They were then shown a brief segment of an article, all participants assigned to a given article topic viewing the same article text. Then they were shown a discussion about the article, and the discussion type varied as an experimental manipulation. As described above, I had ten discussion types, six of which displayed conflict and four of which did not, and an 11^{th} (control) condition where no discussion was shown. I asked comprehension questions to provide greater motivation for participants to read the passages for understanding. The passages were presented as screenshots to exactly maintain Wikipedia formatting and presentation, and prevent participants from using the browser's Find or Copy-Paste features to answer the comprehension questions.

Two pilot versions of this experiment (with about 500 participants each) included an additional no-discussion control condition with extra article text to equalize the amount of time a participant spends on task in the control and the discussion conditions. I found no significant differences between the controls, so I dropped the extended text condition.

Among other questions, I then asked participants to evaluate the article using seven-point Likert items, each with {{Strongly, Moderately, Slightly} {Agree, Disagree}}, Neutral, and "I don't know" options. They evaluated the article with four questions:

"Based on the excerpt shown above, I believe the article as a whole is likely to be…

- … well-written."
- … an accurate and trustworthy source of information."
- … biased." [Scale reversed for analysis.]
- "Based on the excerpt shown above, I believe this article should be included in a collection of high quality articles."

The first three questions used for assessment of perceived article quality are also used in the Wikipedia Article Feedback Tool (Wikimedia Foundation, n.d.) version 4, which asked readers at the end of each article to "rate this page" on a scale of one to five stars on whether the article was "Trustworthy" (tooltip interpretations focused on the page having more or less reputable sources), "Objective" (tooltip interpretations ranged from "heavily biased" to "completely unbiased"), "Complete," and "Well-written." I adopt three of these same measures, omitting "complete" because I was showing participants only a small segment of the article. The wording about inclusion in a collection of quality articles was inspired by the similar dependent measure in (Kittur, Suh, et al., 2008). Wikipedia's assessment tool also includes a checkbox allowing users to identify themselves as "highly knowledgeable" about an article's topic; I include this as a pre-task seven-point Likert item.

Turkers were given the option of participating again so long as I could guarantee that they would not see a discussion type or article topic more than once; in this experiment they were limited by the number of unique article topics (four).

3.7: Study 3: Results

3.7.1: Participants

A total of 1348 surveys were completed by 566 unique Turk workers over the course of 26 hours. I paid $0.60 for each survey and a bonus of $0.15 to the 196 participants who participated the personal maximum of four times. 37 the 1348 surveys, completed by 15 participants, did not meet my *a priori* inclusion criteria that the worker must be in the United States (as discovered by a GeoIP lookup resolving outside the US) and those data were discarded.

Participants in this study were reasonably well-educated, as self-reported on a seven point scale, with 90% reporting at least some college experience, nearly a third reporting a bachelor's degree, and 16% reporting a post-graduate degree. They are also regular Wikipedia users, with over 80% reporting that they read Wikipedia "a few times per week" or more often, and over 96% reading "a few times per month" or more often. Over 85% agreed with the statement "Wikipedia articles are an accurate and trustworthy source of information." All of these factors were randomly distributed across the different experimental conditions.

A free-text box at the end, the only item explicitly labeled "optional," was filled out on 42.7% of surveys, showing that the Turkers were engaged enough to do extra work on the task even though it was not required for payment. 59.6% of the comments were substantive (beyond e.g. "no comments"). Participants reported (in their words) that the task was enjoyable, that it taught some of them to pay attention to their information sources, and that the pay was fair. These survey subsets were not significantly different on rated quality or on how participants thought they had been affected by reading the discussion.

To check for responses that may have been completed too hastily, I manually reviewed all responses that were completed in under three minutes, with that cutoff being slightly above one standard

deviation below the mean. The comprehension question responses and overall response patterns were very similar to the remaining data, so I found no reason to exclude any other participants.

3.7.2: Measures

I examined the four measures of article quality (well-written, unbiased, accurate and trustworthy, and should be included in a collection of high quality articles) to see whether they did indeed all measure a similar construct (overall "quality") or whether they were independent dimensions. I found that they were significantly correlated (average $\rho = .570$; $p < .0005$), and loaded onto a single factor in a Principal Components Analysis, consistent with pilot results. Confident that there was only one underlying dimension, I then considered only the final (overall) quality question, as an interpretable measure of overall perceived quality for my primary dependent variable.

As a manipulation check, I asked participants whether a conflict was present in the discussion. Those in a conflict condition agreed fairly strongly (mean 1.78 Likert points, 95% CI [1.67, 1.88]) and those who weren't disagreed (mean −.65, 95% CI [−.83, −.46]); the difference was highly significant ($p<.0005$). The Avoidance and Competition resolutions were seen as having more conflict than the other conflict types, and referring to policies in removing a source had less, but there was a definite gap between the 95% confidence intervals for the mean measure of conflict in conflict and non-conflict conditions. These data provide evidence that the manipulation was strongly successful.

Of 1200 responses, most agreed with the pre-test statement "Anybody can edit Wikipedia," with the modal response (39%) at "strongly agree;" only 10.3% disagreed with that statement and an additional 4.5% were neutral. However, participants generally did not use Talk pages. When asked how often they read Talk pages, 23.6% said "I don't know what these are" and an additional 44.6% of the 1303 responses said "never." 20.4% read Talk pages "a few times per year" and only 11.4% read them more often.

3.7.3: Analysis

Where means are reported in Likert points, they are on a seven point scale coded for analysis in the range [−3, +3], and otherwise unaltered from the participants' responses. To compare subsets of the data and see if viewers rated the same article text as being of different quality based on the style of discussion they saw, I report the results of a Kruskal-Wallis one-way analysis of variance, which is a nonparametric test that permits ordinal Likert item data. Mean numeric values are reported to allow the reader to estimate effect size.

In order to ensure that my results were valid under strict statistical assumptions regarding fully independent measurements and rule out order effects, I repeated the analysis on just the first survey that each participant loaded, and identically significant effects were observed. I also replicated the analysis using ANOVA instead of Kruskal-Wallis, and again observed the same effects. Treating the data either as ordinal (Kruskal-Wallis) or interval (ANOVA), the results do not vary.

By randomly assigning participants to experimental conditions, I expected prior knowledge and perceptions about Wikipedia, the topic, and the experimental task to be randomly distributed across conditions. I have verified that prior knowledge and trust of Wikipedia are not significantly different between the groups that I am comparing below.

3.7.4: RQ1: How does exposing discussions about article content affect perceived article quality?

My first planned contrast examined whether or not exposure to the discussion (of any type) led readers to assess the article at a different quality level than readers who did not see the discussion behind it. I find that *people who saw any Talk page discussion rated quality significantly lower than those who did not see any discussion* (0.21 vs. 1.08 Likert points, $p < .0005$).

I hypothesize that if people are exposed to conflict about an issue, they would be less likely to perceive a discussion outcome as high quality. I find that *people who saw conflict conditions rated*

quality lower than those who saw non-conflict conditions (−0.04 vs. 0.61 Likert points, p<.0005). I also compared quality ratings of those who saw non-conflict discussions with those in the control condition, and found that *even those who saw non-conflict conditions rated quality lower than those who did not see any discussion* (0.61 vs. 1.08 Likert points, p=.011).

3.7.5: RQ2: Do different kinds of conflict resolution have different effects on perceptions of content quality?

After finding the main differences between no discussion, non-conflict discussion, and conflict discussion conditions, I looked within these groups to see if particular strategies for conflict resolution led to higher or lower quality ratings than others. I found no significant differences between quality ratings among non-conflict discussions; none were expected. The differences among conflict conditions reported in this section provide the basis for the conflict types used in Study 4 below.

Based on the prior literature (e.g. Montoya-Weiss et al., 2001) and pilot data, I hypothesized that the "ignored complainer" in the "avoidance" strategy would lead to significantly lower quality ratings than the other discussion types, because it shows one editor complaining about low quality and nobody responding to those criticisms. Results from this experiment support this hypothesis, whether comparing to just the conflict conditions (−1.13 vs. 0.16 Likert points, p<.0005) or to all discussion conditions (−1.13 vs. 0.36 Likert points, p<.0005).

Based on prior literature that describes collaboration as the conflict resolution strategy generally leading to the best outcomes, I believed the collaboration conflict resolution strategy would be perceived as a "good" conflict resolution strategy and enhance output quality ratings, at least in comparison to other ways of resolving conflicts. I hypothesized that the collaboration condition would lead to significantly higher perceived quality than the other conflict types. Data from the present experiment support this: People who saw the collaboration condition rated quality higher than those who saw other conflict conditions (0.61 vs. −0.18 Likert points, p<.0005). People who saw the

compromise condition also rated quality higher than those who saw other conflict conditions (0.47 vs. −0.13 Likert points, p=.004). The compromise and collaboration conditions were not distinct from each other; they were also illustrated similarly. In fact, these two were not significantly different from the non-conflict discussion conditions in terms of quality rating, even though they were significantly different from the non-conflict discussion conditions in terms of perception of conflict (1.70 vs. −0.65 Likert points, p<.0005). These two resolution strategies were able to overcome the negative effects of conflict on quality ratings.

3.7.6: RQ3: What do participants believe about how viewing the discussion may have changed their perceptions?

Participants who saw discussion conditions were asked directly if reading the discussion affected their perceptions, and if so in what direction. The question, options, and data for each discussion condition are shown in Figure 5. Participants believed that seeing the discussion *raised* their perception of the article's quality and of Wikipedia in general, overall and for many single-discussion-condition subsets, with significant effects indicated by asterisks in Figure 5. The ignored complainer in the "avoidance" condition is the only one where participants were aware of the fact that reading the discussion lowered their perception of that article's quality.

An exact significance test in these analyses examined the balance between the "raised" response options as one set and the "lowered" response options as another set, calculating how likely the actual balance between these would be if both were equally likely, akin to calculating the probability that at least a given percentage of "heads" would appear on the appropriate number of flips of a fair coin. Unless otherwise indicated, significant effects have p<.0005 and non-significant effects have p>.05.

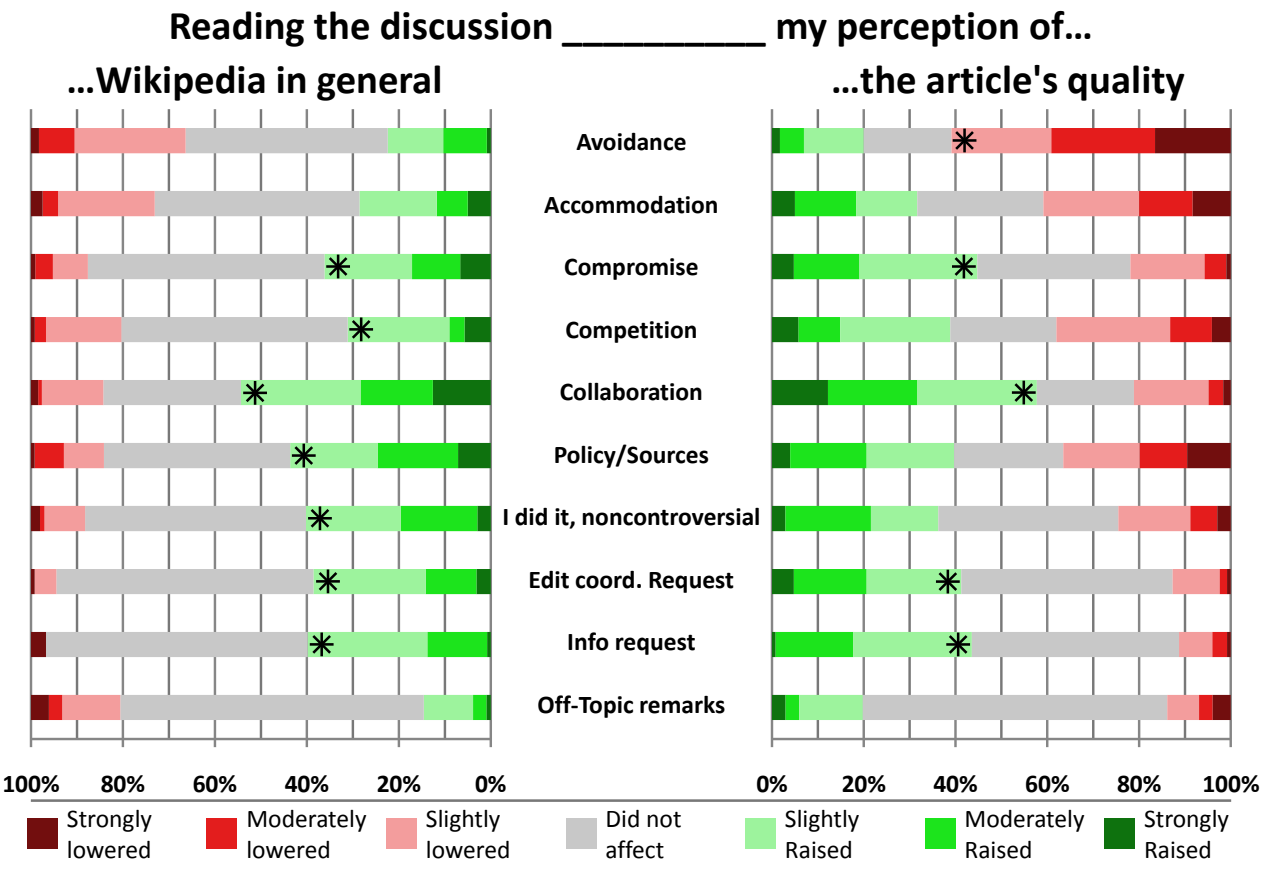

Figure 5: How participants think viewing the discussion affected their perceptions

3.8: Study 3: Conclusions and Discussion

Large-scale distributed systems for collaborative content creation must have a good way of dealing with controversial topics. I have shown that the way these conflicts are dealt with, and whether that is exposed to consumers of the content, impacts how viewers perceive the quality of the content, even when its actual content is held constant.

I find evidence that at least in this setting, surfacing discussions about content lowers the perceived quality of the content, and that this effect differs significantly depending on how the discussion is conducted and whether or not a conflict is revealed. However, the effect runs counter to participants' self-reported perceptions, as participants tended to report that reading the discussion increased their

perception of the article's quality and of the overall platform (here, Wikipedia) in general. The discussion that follows explores possible mechanisms that may help explain the effect.

I suspected that the quality ratings of those already familiar with Talk pages would not be affected by seeing discussion (especially non-conflict discussion) in the same way as somebody who had never seen Talk pages, because the discussion may have made the less Wikipedia-savvy people more aware of the fallibility of Wikipedia's editing process. If the presence of discussion changes how people engage with the article in a way that reduces perceived quality even among individuals already familiar with discussion pages, and the same results are observed in that subset, I would reject that explanation. To test for this, I asked at the beginning of the exercise how often participants read Wikipedia Talk pages, with results summarized above, and filtered to use only responses where people reported familiarity with Talk pages. While some statistical tests suggested the same primary results, when using non-parametric ordinal tests and strictly maintaining the limit of one observation per Turker, I did not have sufficient power to observe significant effects that would lead me to conclusively eliminate this possibility.

This experimental design, across all conditions, aimed to engage participants in reasonably deep thinking. The comprehension questions required all participants to read in the material in a reasonable degree of detail; presentation of those passages as images helped force more detailed reading and engagement.

However, seeing the discussion may have engaged participants' critical thinking skills (Kahneman's System 2) more deeply than the non-discussion condition, causing them to look at the original material more skeptically, leading participants to be more critical of the article in their ratings. Interestingly, they nevertheless believed the discussion caused them to perceive the article as higher quality, perhaps because they believe they were able to make a more accurate and informed evaluation, or because

System 2 included a more thoughtful understanding of "quality" than the quick intuitive judgment of System 1, which often substitutes easier questions that may have different baseline answers (Kahneman, 2011). If this explains the results, it would suggest that designers of online collaborative content creation systems need to pay close attention to factors that influence System 2 engagement (summarized in Kahneman, 2011) when considering how people will perceive content quality.

As another possible explanation explored in Study 4 below, design literature indicates that people are less willing to be critical of work that they perceive to be "finished" or "complete," and more willing to offer criticism of works in progress. People are more willing to give higher-level constructive criticism about something that seems to be more "sketchy" with lower level details not yet fixed, as compared to a polished product (Y. Y. Wong, 1992). This may be a mechanism of System 1.

All article segments used in this study were taken from works currently listed as Featured Articles, meaning they have already reached the highest quality standard on Wikipedia, but all Wikipedia articles are in some sense incomplete (Wikipedia community consensus, 2012b). Revealing the discussion might frame the article more as a "work in progress" than a completed, polished piece. This more apparent state of incompleteness could invite more criticism, as reflected in lower quality ratings, even while participants feel that reading the discussion improved their perceptions (presumably compared to other work viewed as being in a similar state of completion). This can be investigated in a future study that manipulates the perceived completeness of collaboratively produced content, regardless of the presence or absence of discussion. If this explains the experimental results, system designers would need to attend to stylistic details that make work seem more or less polished, depending on their goals for the system. In light of the findings about how different types of conflict resolution strategies impact perceptions of quality, designers of systems that expose discussion with content might also consider ways of detecting and classifying those strategies to proactively flag certain future discussions for moderator attention and possible intervention.

This study includes an experiment and a robust set of results about *how* the presence of discussion causally affects perceived quality. In this work, I have found a "sausage" effect that revealing any discussion can lower perceived article quality, with the strength of the effect depending on the presence of conflict and the way that any present conflict is resolved.

From this study, we see that readers in an online deliberation platform can be affected by design decisions concerning what discussion is exposed alongside content, and what that discussion contains. How far does this effect extend? Study 4 explores the limits of this effect with different materials (which are framed as proposals, rather than being mentally modeled as reference works) and reveals that even when readers are affected by discussion about content, they may be able to evaluate additional content even by the same author or about the same mix of topics independently of that effect.

Study 4: Conflict in Comments: Learning and the Limits of Carry-Over

4.1: Study 4: Chapter Summary

Study 3 above suggests that when exposed to discussion related to a particular piece of crowdsourced text content, readers may perceive that content to be of lower quality than readers who do not see those comments, and that the effect is stronger if the comments display conflict. This study presents a controlled experiment with over 1000 participants testing to see if this effect carries over to other documents from the same platform, including those with similar content or by the same author. Although I do generally find that perceived quality of the commented-on document is affected, effects do not carry over to the second item and readers are able to judge the second in isolation from the comment on the first. I confirm the prior finding about the negative effects conflict can have on perceived quality but note that readers report learning more from constructive conflict comments.

4.2: Study 4: Introduction

In many platforms designed to support large-scale distributed problem solving, basic ideas are crowdsourced and posted publicly, where others can evaluate the contribution and/or add comments and/or read the comments others have written. However, there are questions being raised in at least the public dialogue about the potential of spill-over effects from comments and interactions posted online to undermine the significant positive potential of Internet-based collaboration (Stein, 2016).

Where others' comments on or ratings of content are also visible, as is often the case, they can have an impact on the level of quality perceived by subsequent readers. Social influence on perceptions of quality could create a feedback loop that leads groups of people to irrationally herd toward large group evaluations that are path-dependent and not necessarily connected to the quality of what is being perceived (Bikhchandani, Hirshleifer, & Welch, 1992; Salganik et al., 2006). Causes and consequences of market-based valuation bubbles and herding in asset pricing have been studied (e.g. Hirshleifer & Hong Teoh, 2003), but little is known about how much of that knowledge applies to distortions in perceptions

of value when the signals and consequences are more purely social, in the absence of pricing or direct economic incentives.

Examining social influence effects in large online community platforms, Muchnik, Aral, & Taylor conducted an experiment in which fresh and unrated content, which users had posted to a social news aggregation Web site, was *randomly assigned* to an initial upvote or downvote (or control condition with neither). The random upvote increased the probability of up-voting by the first viewer by 32% without a corrective increase in down-voting; the effects of that initial manipulation persisted and increased mean ratings by 25% five months later. A random initial downvote doubled the probability of subsequent downvotes, but also significantly increased the probability of a subsequent corrective upvote. "Friends" of the commenter were more likely to upvote in response to a randomly assigned initial vote (up or down) than to a post with no random initial vote, but even when differences in the probability of voting were considered, there remained differences attributed to statistically significant opinion change resulting from the random initial vote. Their results "suggest that social influence substantially biases rating dynamics in systems designed to harness collective intelligence." That paper concludes by calling for more research exploring mechanisms driving individual and aggregate ratings, especially in real social environments, as "essential to our ability to interpret collective judgment accurately and to avoid social influence bias in collective intelligence" (2013).

In the setting of Wikipedia, Study 3 above showed that when readers are exposed to discussion behind collaboratively edited content, their assessment of the quality of the content is lower than readers who do not see the discussion. The effect is stronger if the discussion contains conflict, but this strengthening can be erased if the discussion shows a Compromise or Collaboration resolution strategy from the editors involved. The effect, observed in a between-subjects experiment, was in that study counterintuitively accompanied by participants' self-reports that reading the discussion *increased* their perceptions of article quality.

The present study seeks to test the extent of that finding, using a similar methodology in a different setting and measuring the extent to which this effect might extend beyond the article being discussed directly, to other articles that are similar along a couple selected dimensions.

4.3: Study 4: Method Overview

In this study, I showed participants a segment from each of two crowdsourced proposals entered into a platform designed to support large-scale collaboration around a complex issue, namely global climate change. Following the first proposal (only), most participants saw a comment associated with that first proposal. I experimentally varied the type of comment shown in a 3x2 full factorial design, plus a no-comment control, as described in "4.5.2: Experimental Conditions" below. I also independently varied the relationship between the first and second proposals (i.e. same author, similar topics, or neither) as described below. After reading each proposal (and if applicable, the comment immediately following the first proposal), participants were asked to evaluate the proposal quality with a multi-item scale. I tested to see if those quality evaluations changed as a result of the experimental conditions to which participants were randomly assigned.

4.4: Study 4: Hypotheses

Hypotheses in this work are presented here in descriptive form, summarized in Table VI. Based on work in Study 3 above, I hypothesize that (H1) the presence of a comment will negatively affect participants' ratings of the first proposal's quality, and (H2) the size of the effect will depend on the type of comment, such as whether it presents conflict. The size of the effect may also depend on how that conflict is resolved, as in Study 3. This study investigates further to determine if (H3) the effect extends to other proposals (i.e. the second one participants saw) that are similar or different in certain ways.

In investigating that possible spill-over effect, I experimentally manipulate whether comments are directed at the *content* or the *author.* I hypothesize that (H4) if the comment is attributed to something about the *content* (e.g. a claim that the main idea is significantly flawed), the effects of reading that

comment might extend to other topically similar proposals regardless of who wrote them, and (H5) that if the comment is something attributable to the *author*, the effects might extend to other proposals by the same author regardless of proposal topic.

Whether or not H1 and H2 regarding the effects of comments on the rated quality of the first proposal were supported, I wanted to be able to better understand the processes underlying any differences in ratings between experimental conditions. If the same set of processes is at work here as observed in Study 3 above, I hypothesize (H6) that participants would report *beliefs* that their perceptions of quality had been affected in the *opposite* direction than the between-subjects analysis showed that it had been. To explore this hypothesis, I asked participants who saw comments about the degree to which they thought reading the comments raised/lowered their quality perceptions, as two questions on a page immediately following participants' evaluations.

I also asked participants to indicate whether or not they believe they learned anything from the comments, to investigate a potential benefit they might provide and put H2 (regarding the effect of conflict on perceived quality) in context, as readers might believe that comments offering a different perspective provide them with more novel and/or useful information than comments that do not express both a conflict and resolution. This allowed me to explore the hypothesis (H7) that participants would at least self-report learning most from comments containing constructive conflict.

In all conditions, content was presented in the same order. I recognize that reading the first proposal could have an order-specific effect on the rating of the second, separate from underlying differences between the proposals that might lead to rating differences between the first and second proposals. This experiment specifically investigates one hypothesized order effect, namely how reading a comment and then providing an evaluation that may be influenced by it may also influence perceptions of material read immediately after it, on the same page. Having the rating materials on the

same page increases the probability of observing (and being better able to characterize) a carry-over effect if one exists. Participants are randomly assigned to a comment condition (or control condition absent any comments), and the primary analysis compares between those conditions.

4.5: Study 4: Materials

4.5.1: Corpus

Inspired by examples of successful large-scale collaboration elsewhere, such as Wikipedia or FoldIt, Malone et al. created the MIT Climate CoLab to be a global platform for collaboratively developing and evaluating proposals for what to do about global climate change (Malone et al., 2014). In an annual series of contests, its members have collectively produced, commented on, and voted on over 1500 proposals, typically 1500-3000 words in length (Climate CoLab, n.d.-c), on a wide range of climate-change-related topics.

In recent years, several contests have been run in parallel, each addressing different subtopics (such as Transportation, Waste Management, Buildings, and Energy Supply) or calling for plans that integrate a number of other proposals (Malone et al., 2017). Participants and problems discussed come from many countries, though a majority of participants come from North America (Duhaime et al., 2015b). The interface, proposals meeting the language requirements (Carey, 2015), and all materials used in this experiment are in English. At the start of the main 2016 contests, the site had over 50,000 registered user accounts and 1501 completed proposals, which served as the base set of documents in my model and experiment.

Materials were further restricted to the 350 Climate CoLab proposals submitted in contests between 2011 and 2015 inclusive that made the semifinalist round as selected by expert judges. I chose semifinalist proposals (or higher) to establish a minimum degree of quality, similar to the procedure for choosing high-quality articles used in Study 3 above. I set the threshold at semifinalist proposals as

opposed to finalists or winners to maintain some quality while also ensuring that the pool of proposals to choose from would be sufficiently large (as in Ozturk, Han, Towne, & Nickerson, 2016).

I further eliminated proposals that could not be basically understood by the summary text, e.g. because the proposal relied heavily on reference to other proposals or documents. I removed two proposals whose content had been submitted in the form of images, and another whose latest version simply read "The new version will be posted shortly" because the text of these had not been properly captured for topic modeling.

4.5.2: Experimental Conditions

Comments shown on the first proposal were randomly assigned from a 2x3 factorial design, plus a "no comment" control which helps test H1. The three levels of the "conflict" factor (Conflict without resolution, Conflict with resolution, and Non-conflict) were chosen based on the categories distinguished by Study 3, and help test H2. The two levels of the "direction" factor (content-directed vs. author-directed) help test H4 and H5.

The second proposal was chosen randomly from the cells in Table V. I excluded the "Same author, similar content" cell because when observed, the proposal pairs in this cell are largely copies of each other, sometimes with little or no modification, and in allocating budget for experimental conditions, it was least interesting to see if comments on one proposal would affect perceptions of another copy of that proposal. The relationship between proposals was called out in bold underlined header text just before the presentation of the second proposal, as shown in Figure 6.

4 similarity types (2x2):	*Same **A**uthor*	*__D__ifferent Author*
*Similar **C**ontent*	≈ Copies (not tested)	Relationship C
***D**issimilar Content*	Relationship A	Relationship D

Table V: Relationship second proposal has to first proposal, randomly assigned.
This factor crosses author and content similarity factors, a factorial design excluding one cell.

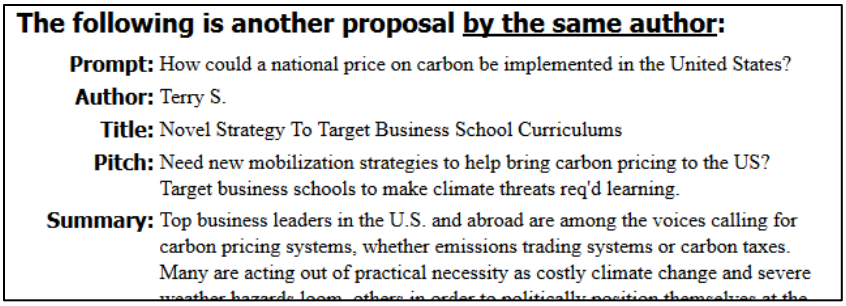

Figure 6: The relationship between proposals was explicitly called out. This screenshot shows Relationship A from Table V. For Relationship C, the underlined text read "about a similar topic" and for Relationship D, it read "that was submitted" without underline.

4.5.3: Selecting focal proposals

As potential focal proposals (those shown first in the experiment), I considered those where other proposals existed that fit the "same author, unrelated content" criteria described below. I selected multiple focal proposals so that the experimental results would not be too sensitive to specific details of any particular proposal, but kept the number small so that I could collect relatively robust sets of quality ratings on each and analyze aggregated results while still controlling for any real quality differences that may exist between selected proposals.

In order to better cover the space of available proposals, I used a constraint satisfaction solver to maximize diversity by selecting the set of four focal proposals by four different authors that were on average maximally different from each other according to the LDA/cosine similarity measure used in Study 2 above (also used in "4.5.4: Selecting topically related proposals" below). The solver I used was Excel 2013's "Evolutionary" solver, which produced better results than its "GRG nonlinear" solver.[12]

4.5.4: Selecting topically related proposals

For each focal proposal, I selected the proposal that had no overlap in author team and the greatest topical similarity to the focal proposal, operationalizing relationship "C" from Table V. Topical similarity

[12] For more detailed citations, see http://www.solver.com/ excel-solver-algorithms-and-methods-used.

was measured by the cosine similarity of proposals' topic vectors according to a Latent Dirichlet Allocation (LDA) model run over the entire corpus with program default hyperparameters.

Consistent with Study 2 above and with Xu and Ma's goal of maximizing dissimilarity between clusters (2006, p. 303), I selected the number of topics by choosing the model with the lowest average percentage of proposals that has neither or both topics in each possible topic pair. The model was run to 1000 iterations for coarse tuning between 5 and 300 topics, and 2000 iterations for fine tuning between 50 and 60 topics, concluding with a 57-topic model, as the one which maximally separated proposals into different topics. Using this measure of topical similarity instead of the CoLab contest categories helps make the results of the study more general than the Climate CoLab structure which requires manual creation of a topic hierarchy and manual assignment of proposals to those categories (here, assignment is done by proposal authors who are not experts about the categorization scheme).

4.5.5: Selecting topically unrelated proposals

For each focal proposal, I sorted other proposals according to the same LDA/cosine similarity scale and randomly selected from the bottom half of that distribution, ensuring that it was by a different author, operationalizing relationship "D" referenced in Table V.

Proposals operationalizing relationship "A" in Table V were drawn from this same bottom half of the distribution. If the author had multiple proposals there, I chose the one more different from the focal proposal (this happened in only one of the four sets, and the two options were adjacent to each other in the list). In all eight cases, these proposals were from contests different than the focal proposal.

4.5.6: Proposal Tweaks

Proposal summaries were modified from the originals for greater suitability for use in the experimental setting and consistency with each other in the following ways. Hyperlinks and text references to other proposals were removed, to make the summary more self-contained. Special

formatting was removed. Acronyms that were not clearly introduced but discoverable in the original context (e.g. GHG) were spelled out. Spelling and grammar were cleaned up so that these attributes of writing would not dominate or interfere with perceptions of quality based on other factors, removing this potential source of variance for a cleaner experiment. One proposal summary was shortened (largely by removing redundancy) to bring its length in line with the others, appropriate for a Mechanical Turk task. After these modifications, the proposal summary lengths ranged from 168 to 350 words, with an average of 290 and standard deviation of 48. Material length was comparable with a previous study based on news articles (Steinfeld, Samuel-Azran, & Lev-On, 2016).

4.5.7: Selecting names

Because author attribution is a factor being explored in this experiment, author names were attached to proposals to help underscore the "same author" relationship. However, researchers presenting experimental materials with names attached need to be aware of the impact of those names, because aspects of the name (such as perceived gender) can change the way participants perceive the materials attributed to them (Fleet & Atwater, 1997).

Through a set of four studies, Fleet and Atwater (1997) identify four most gender-neutral *first/given names* and of these four I chose the two most prevalent ("Expected total number alive today" male + female total) in the United States according to the Wolfram|Alpha Knowledgebase, 2016 (e.g. Wolfram Alpha, 2016), which were Terry and Lee respectively.

I used only a surname initial instead of a full surname according to the advice of Kasof (1993, p. 152). The US Census publishes the frequency of *surnames* occurring >100 times, covering 90% of the population (Word, Coleman, Nunziata, & Kominski, 2016); at the time of study the most recent available data was from the 2000 Census. (First name information is not available from the 2000 census.) Within this data, the most common first letter (surname initial) is M, followed closely by S.

I crossed these two for balance. In this experiment, the focal proposal was always credited to "Terry S." and the other proposal, if by a different author, was credited to "Lee M." An experiment by Howard & Kerin (2011) found that *similarity* between a participant's name and the putative name of an author whose work the participant was evaluating (in the experiment, first name and last initial were shown to match or not match participants'), engages self-referencing and increases thoughtful examination. Here, I did not ask for participants' real names, but I don't expect any rare overlaps to lead to differences between the randomly assigned experimental groups.

4.5.8: Dependent Measures

As in Study 3 above, I consider quality assessment criteria used in the community where the data comes from. In the Climate CoLab, proposals are assessed by expert judges along four scales: Feasibility (in four specific aspects), Novelty, Impact, and Presentation (Climate CoLab, 2016). These criteria and their descriptions have been consistent through the history of the platform to the time of study. Each participant was asked to evaluate each proposal using the following seven-point Likert items, each with {{Strongly, Moderately, Slightly} {Agree, Disagree}}, Neutral, and "I don't know" options:

"Based on the excerpt shown above, I believe the proposal as a whole is likely to…

…be {technically, economically, socially, politically} feasible." (4 questions, answers averaged for feasibility)

…be novel, reflecting innovative thinking and originality."

…make an impact on the issue raised in the prompt, if it were implemented."

…be well-presented."

For an overall measure of quality, on the same scale, I also included an item (from (Kittur, Suh, et al., 2008) and Study 3 above) *"Based on the excerpt shown above, I believe this proposal should be included in a collection of high quality proposals."* This is the fifth item in a five-item perceived quality scale

which is my primary dependent variable. In relatively rare cases where a participant refused to answer any particular item, that item simply carries no weight in computation of the composite (mean) rating for that participant and scale.

4.5.9: Demographics & prior expertise

Before the task, participants answered questions about their educational background and level of interest and experience in the general domain area of the content material, including questions copied from prior surveys of the community where materials came from.

Prior knowledge about a topic has been identified as a potentially dominant factor in quality evaluations (Steinfeld et al., 2016) in work that explicitly seeks follow-up studies illuminating the impact of issue familiarity on the effect of comments (Steinfeld et al., 2016, p. 71). Also, people who have more prior knowledge or are more involved do more elaborate information processing, attend to more quality cues, especially intrinsic ones, and may be less extreme but possibly faster in their overall quality judgments (Steenkamp, 1990, p. 315). Therefore, after each evaluation, I included an item "I have expertise on the topic(s) discussed in this proposal." As the final multiple-choice question in this study, I asked participants about their level of familiarity with the platform from which materials were drawn.

4.5.10: Participants & Filters

I recruited participants on Amazon's Mechanical Turk paid crowdsourcing platform, in part because the population for generalizability of results is intended to broadly include those likely to visit/use crowdsourcing platforms. (Materials in this experiment come from a different crowdsourcing platform, described above.) This choice of participant pool also has several other benefits relevant to experiments exploring the impact of social information on perceptions in e.g. *CHI* (Hullman, Adar, & Shah, 2011). Experimental tasks were posted midweek, on days described as slow (with respect to the volume of tasks being posted) on the Turker Nation forum. Tasks were posted through afternoon and evening times and continued at a slower pace into the following day.

I reduced potential evaluation variance by restricting participants to those who were in the United States according to their Turk profile, and prior to analysis I filtered out any data from participants who did not have a GeoIP lookup resolving to the United States. I also restricted participation to those who had at least 500 assignments approved by other requesters and a 95% overall approval rating. (This is the same as in Study 2 above). In the main experiment, I also excluded from analysis two participants who wrote keyboard-mashing strings in unselected "other" boxes on demographic questions, and two participants who took steps to defeat the participation limits. I filtered out those who completed the main experiment in under 2.5 minutes because less than one minute per proposal plus 30 seconds for demographic questions indicates those participants are less likely to have fully read the materials or questions.

Both usage goals and time pressure also affect perceptions of quality (Steenkamp, 1990, p. 316), and I held these constant across experimental conditions. During the task, I specified usage goals (proposal evaluation) and did not add any time pressure beyond what is self-imposed by participants independently of the task (similar to e.g. Steinfeld et al., 2016), permitting an hour for a task that typically took several minutes.

At the very end of the experimental task, participants also had an open text box to optionally provide feedback about what they had just completed. I have found this to be a good practice which can facilitate detection of certain errors in experimental setup if they exist, provide qualitative feedback that can help inform future analyses and/or task designs, and increase participant satisfaction by allowing participants to express any remaining thoughts they wanted to express. I read all feedback submitted.

4.6: Study 4: Manipulation Check on Comments

4.6.1: Manipulation Check on Comments: Design

I created the comments for each focal proposal and experimental type based on a review of real comments left on proposals on the site, similar to the setup in Study 3 above. Before proceeding with the experiment, I checked to see if the differences between experimental conditions were manipulating the intended constructs. To do this, each of the 24 comments was posted to Turk along with the following 8 statements assessed on the same 7-point agree/disagree scale described above, followed by a free-text feedback field:

1. This comment is civil.

2. This comment is directed at the *author* of the proposal.

3. This comment is directed at the *content* of the proposal.

4. The author of the comment and the author of the proposal likely have some conflicting views, at least regarding what this comment is about.

5. If there is conflict present, it seems likely to have a good resolution. (If there is not conflict present, please choose "neutral.")

6. If this comment were automatically analyzed, it **should** be scored as a *negative* comment.

7. If this comment were automatically analyzed, it **should** be scored as a *positive* comment.

8. This comment is likely to be helpful to the person who wrote the proposal the comment is on.

I paid US$0.12 per comment reviewed. In this manipulation check, participants were allowed to participate up to 6 times, each time randomly selecting one of the six comment types not previously evaluated by that worker, and randomly selecting one of the four focal proposals. Participants in the manipulation check were not allowed to participate in the main experiment.

4.6.2: Manipulation Check on Comments: Results

160 unique Turkers, who passed the same inclusion restrictions applying to the main experiment, completed 432 rating tasks. Based in part on the results detailed below, I believe the comment type manipulations were successful.

4.6.2.1: Comparison between (focal) proposals

I checked each of the eight questions, plus time on task, and (using an ANOVA) looked for any significant differences between the four focal proposals. In cases where the Levene statistic rejects the null hypothesis that group variances are homogeneous, I use the Welch statistic as it is more appropriate than the *F* statistic to test for the equality of group means. I found only 3 significant differences (α=.05), which matches the expected number of randomly significant comparisons with 6 comment types and 9 comparisons each (6*9=54 comparisons). When all comment types were aggregated together, there were no significant differences between focal proposals on any of the 9 measures.

4.6.2.2: Author-directed vs. Content-directed: Questions 2 & 3

The author-directedness of the comment (Question 2) was significantly ($p<.0005$) higher in the author-directed conditions, considering the data overall (4.09 vs. 6.22 Likert points) or each proposal set independently, or each level of the "conflict" factor independently.

The content-directedness of the comment (Question 3) was significantly ($p<.0005$) higher in the content-directed conditions, considering the data overall (4.63 vs. 6.50 Likert points) or each proposal set independently, or each level of the "conflict" factor independently.

4.6.2.3: Conflict vs. non-conflict conditions: Questions 4 & 5

The comment's level of conflict (Question 4) was significantly ($p<.0005$) higher in the conflict conditions, considering the data overall (2.49 vs 5.53 Likert points) or each proposal set independently, or each level of the "direction" factor independently. The 95% confidence intervals for the mean of this

item are fully on the "disagree" side for non-conflict conditions and on the "agree" side for conflict conditions.

Excluding non-conflict comment conditions from analysis, the conflict's likelihood of a good resolution (Question 5) was significantly ($p<.0005$) higher in the constructive conflict conditions than the "no good resolution" conditions, considering the data overall (3.16 vs 5.01 Likert points) or each proposal set independently, or each level of the "direction" factor independently ($p=.001$ in content-directed conditions).

4.6.2.4: Timing

The amount of time taken to complete the rating task was not significantly different across comment types, considering the data overall or each proposal set independently.

4.6.2.5: Comment type 4 more negative

The civility, positivity, and helpfulness of the comment (Questions 1, 7, & 8) were significantly ($p<.0005$) lower, and negativity (Question 6) significantly higher, in comment type 4 (author-directed unresolved conflict) than in the other comment types, considering the data overall or each proposal set independently.

4.6.2.6: Sentiment and Civility: Questions 1, 6, 7 & 8

All four of these questions were significantly correlated with each other ($p<.0005$ by Pearson or Spearman). Questions 6 and 7 were very strongly negatively correlated (Pearson's r: -.937; Spearman's ρ: -.930), as expected.

Comments were rated significantly ($p<.0005$) less civil, less positive, and more negative in conflict conditions than in non-conflict conditions, considering the data overall or each proposal set independently, or each level of the "direction" factor independently. Comments were rated significantly ($p<.0005$) less civil, less positive, and more negative in the conflict than non-conflict conditions even

when excluding comment type 4 from analysis, overall or considering each level of the "direction" factor independently.

There are no civility or negativity differences between the two non-conflict conditions analyzing the data overall; the author-directed non-conflict condition is significantly ($p=.001$) more positive than the content-directed non-conflict condition.

In general, comments were scored as significantly ($p<.0005$) more civil and helpful when directed at the content than when directed at the author, overall and (with $p<.05$ on "civil") considering each proposal set independently.

There were no significant differences between author-directed and content-directed comments in positivity or negativity, considering the data overall.

The helpfulness of the comment (Question 8) was significantly ($.0005<p<.01$) lower in conflict than non-conflict conditions when analyzed overall.

4.7: Study 4: Main Experiment Results

1252 responses remained after filtering as described above. The median task completion time was 331.5 seconds, with an interquartile range of (246, 463) seconds.

4.7.1: Multi-Item Scales

In this subsection, I describe and validate the multi-item scales used in my primary dependent measure. I provide evidence supporting the unidimensional treatment of my dependent construct "rated (or 'perceived') quality."

Cronbach's Alpha for the four-item feasibility scale is .869 (n=2391). An exploratory Principal Components analysis with these four items also showed all four loading onto a single principal component, and all four items were significantly ($p<.0005$) and strongly (average >0.6 by Pearson or Spearman) correlated with each other.

Cronbach's Alpha for the five-item quality scale, which includes the feasibility scale as one item, is .865 (n=2343). An exploratory Principal Components analysis with these five items also showed all five loading onto a single principal component, and all five items were significantly (p<.0005) and strongly (average >0.55 by Pearson or Spearman) correlated with each other. This is my multi-item perceived quality scale.

In order to increase comparability across the four proposal sets and reduce noise due to inherent differences in proposal quality, I compute the overall average quality rating for each proposal and compute each participant's rating as a deviation from that proposal-specific mean. This centered measure is my primary dependent variable for proposal quality evaluations discussed below. Unless otherwise noted, reported results are from planned contrast analyses within a larger ANOVA across comment types, with or without the assumption of homogeneous variances.

4.7.2: Main differences in ratings of focal proposal

Differences in the type of comment shown on the focal proposal caused significant differences in how participants rated the quality of that proposal. The conflict conditions led to significantly lower ratings of quality than the non-conflict or non-comment conditions (-.3071 vs .4334 Likert points, p<.0005), replicating the negative effect of conflict observed in Study 3 above and **supporting H2**. (Table VI below summarizes findings for each hypothesis.)

I did not observe any significant differences in the ratings of either proposal based on conflict resolution within conflict conditions, nor between "non conflict" and "no comment" conditions. Finding no significant difference in the latter comparison means I **do not find support for H1**.

4.7.3: Differences in ratings of second proposal

There was a small but significant (p<.0005, n=1249) correlation between participant's quality ratings of the first and second proposal (r=.229, ρ=.194 for deviation from proposal means as used here; r=.188,

ρ=.155 for raw scale score), most likely indicating a per-rater effect that some raters were generally more generous and others more strict. My experimental design controls for this by randomly assigning participants to experimental conditions, so I would not expect any systematic differences in rater scoring bias between the experimental conditions.

Before adjusting for proposal-specific averages, the proposals presented second were rated slightly but significantly lower (4.6156 vs 4.8348 Likert points, p<.0005), which may be due to actual quality differences between the proposals. After participants' ratings are adjusted relative to the average for that proposal (as is done in the primary dependent variable for this study), the average differences in ratings *given by the same participant* for the first and second proposal they saw is not statistically different from zero (95% confidence interval is [-.0858,+.0847]).

I did not find any statistically significant differences in the ratings of the second proposal caused by differences in the type of comment displayed on the first. Although differences in the rating of the first proposal based on comment type were observed, I did not find comment-caused differences in ratings of the second proposal even when the second proposal was actually and was labeled as being by the same author or about similar topics. This is a **lack of support for H3, H4, and H5**.

4.7.4: Learning from comments

Most participants (>60% at each level of the "conflict" factor) agreed with the statement "I learned something from the comments." Participants reported strongest agreement (along the same 7-point Likert scale described above) with "I learned something from the comment(s)" in the "constructive conflict" comments, followed by the "non-conflict" and then "unresolved conflict" comments (group means of 5.18, 4.85, and 4.57 Likert points respectively, all differences statistically significant even with Tukey's Honestly Significant Differences test). This provides **support for H7.**

Further, participants' self-reported learning from the comments correlates significantly ($p<.0005$ overall and for each conflict type) with how much they thought "reading the comments *raised* my perception of proposal quality" ($r=.567$, $\rho=.550$, $n=1044$ overall). The correlation was strongest ($r=.768$, $\rho=.751$, $n=333$) in the non-conflict condition.

As expected, the degrees to which participants thought the comments *raised* and *lowered* their perception of proposal quality were significantly ($p<.0005$ in each conflict condition and overall, $n=1053$ overall) negatively correlated with each other ($r=-.427$; $\rho=-.449$ overall). The correlation was strongest in the constructive conflict condition ($r=-.689$; $\rho=-.697$, $n=375$). In general, significant differences on one or both of these measures were observed alongside significant differences in actual quality ratings, in a direction *consistent* with the observed effect, indicating **no support for H6**.

At least half the participants (at any level of the "conflict" factor) believed that the comments did not lower their perception of proposal quality (i.e. selected a "disagree" option on the "lowered" question) and most participants believed that reading the comments raised their perception of proposal quality. This aligns with the finding from Study 3 that participants believe seeing discussion increases their perception of the quality of the material discussed.

4.7.5: Power

Where I do not observe a difference between experimental groups, that logically means that either (A) no significant difference exists between the groups, (B) the experiment was not designed properly to detect a difference, or (C) the difference between groups was so small that this experiment did not have sufficient power to detect it. (B) seems relatively unlikely given how closely the experimental design mirrors that of Study 3 above, which had previously identified differences, as well as the results of the manipulation check above. (C) appears relatively unlikely, based on computations using G*Power 3.1.9.2 (Faul et al., 2009) to identify how small of an effect I would have still had a 95% chance of

observing, if it existed. For the ANOVA used here, the effect size parameter *f* is considered "small" at .1, "medium" at .25, and "large" at .40 (Cohen, 1969, p. 348).

Aggregating across relationships between proposals (Table V conditions) and looking for differences between second proposal ratings caused by comment types, I would have had a 95% chance of observing differences between cells with an effect size index of .13 or greater, and .10 or greater in a two-group comparison (e.g. conflict conditions vs. others). The minimum effect size observable with 95% probability is .17-.18 for a two-way comparison within any one of the Table V cells. The power for observing a medium size effect is 99.9% for a two-group comparison between comment types within any one of the Table V cells and higher for even a 7-group comparison aggregating across them.

4.7.6: Demographics

Prior work demonstrates that pre-existing cultural context can impact the way readers are affected by comments (Steinfeld et al., 2016). I report demographics here for easier comparison to other participant populations and more detailed identification of the participant pool. The initial questions and responses were as follows, except that questions 2 and 6 also each had an infrequently used "other" option manually recoded for analysis into the most similar available listed category.

1. Do you think that global warming is happening? {Yes: 85.7%, No: 6.5%, Don't know: 7.8%, n=1237}

2. Assuming global warming is happening, do you think it is … {Caused mostly by human activities: 49.8%, Caused mostly by natural changes in the environment: 10.3%, None of the above because global warming isn't happening: 2.7%, Caused by both human activities and natural changes: 37.1%, n=1250}

3. What is your gender? {Female: 50.4%, Male: 49.6%, n=1241}

4. What is your age? {under 20: 1.4%, 21- 29: 31.0%, 30- 39: 33.9%, 40- 49: 18.3%, 50- 64: 13.3%, 65 and over: 2.1%, n=1249}

5. What is the highest level of education you have attained to date? {High school or less: 17.1%, Attending college / university: 20.6%, Graduated from college / university: 46.3%, Attending graduate or professional school: 2.8%, Completed graduate or professional school: 13.3%, n=1245}

6. What is your current situation / status? {Student: 7.6%, Employed full-time: 54.7%, Employed part-time: 14.0%, Free-lance consultant: 7.5%, Unemployed [or homemaker]: 12.9%, Retired: 3.3%, n= 1244}

Last (after task). I am familiar with the MIT Climate CoLab. {Strongly disagree: 62.8%, Moderately disagree: 21.2%, Slightly disagree: 6.6%, Neutral 4.5%, Slightly agree: 2.9%, Moderately agree 1.1%, Strongly agree 0.9%, n= 1234}

4.7.7: Findings with Demographic Covariates

In addition to primary hypotheses looking at how randomly assigned experimental conditions affect the target variables, I also did some analyses exploring how some of my observations may have interacted with demographic variables, which may help guide future research.

Among participants with no college degree, the constructive conflict led to significantly *lower* ratings than the unresolved conflict (contrast value .3865 points, p=.006-.008), while college graduates tended to rate the proposal quality significantly higher after viewing a non-conflict comment than after viewing no comment (contrast value .3141 points, p=.019-.021). Among participants reading comments illustrating unresolved conflict, college graduates rated the proposal slightly higher than non-college grads (contrast value .2950 points, p=.038), even though in the no-comment control condition college grads on average rated proposals slightly lower (contrast value .4044 points, p=.016-.022). This suggests that the more educated readers may have viewed the materials with a bit more critical and independent thinking.

Among comment conditions, the author-directed comments surprisingly led to higher quality evaluations than the content-directed comments in the "Constructive Conflict" conditions only, (-.4843 vs -.2048 Likert points, p=.017), which was enough to create an overall difference between content- and author-directed conditions (-.1564 vs. .0365 Likert points, p=.009). In further exploration, this significant difference in rating was only observed among female participants but not male participants, and among college graduates but not those without a college degree. Rater gender and college education were independent according to a chi-square test with a 95% chance of observing a "small" effect (effect size index w=.10 (Faul et al., 2009)).

4.7.8: Effect of Expertise

I found a small but statistically significant correlation between self-reported expertise on a proposal's topics and participants' quality rating of that proposal, both for the first proposal (r=.078 / p=.006; ρ=.093 / p = .001, n=1242) and the second proposal (r=.103 / p<.0005; ρ=.070 / p = .013, n=1250). There was also a small correlation between self-reported expertise on the first proposal's topics and if the comments raised participants' quality perceptions (r=.142; ρ=.133, p < .0005, n=1055). In general, more than three quarters of participants' self-reported topic-specific expertise level was on the "disagree" side of the scale and less than 14% of expertise self-reports were on the "agree" side of the scale (n=2492). Especially combined with random assignment to the experimental conditions, I do not believe that pre-existing judgments about the material dominated quality evaluations (as when different materials were used by Steinfeld et al. (2016)).

4.8: Study 4: Conclusions

#	Hypothesis
~~H1~~	Comment presence lowers 1st proposal qual. ratings
H2	Comment type & conflict affects size of H1 effect
~~H3~~	H1 effect extends to other proposals from platform
~~H4~~	Content-directed comments on one proposal affect quality ratings on a topically related proposal

#	Hypothesis
~~H5~~	Author-directed comments on one proposal affect quality ratings on another one by that author
~~H6~~	Participants report perceptions being affected by comments in opposite direction from true effect
H7	Participants learn most from comments containing constructive conflict

Table VI: Summary of hypotheses. "No support" results where power calculations indicate a true effect would likely have been seen are indicated by ~~red strikethrough~~ on the left. Hypotheses supported by statistically significant differences are indicated by green underline in the left column.

In this large-scale (Hullman et al., 2011) experiment, I showed that when comments about crowdsourced content are presented alongside that content, the contents of the comment affect how people perceive the content. I replicated a previous finding that comments containing conflict lower perceptions of content quality more strongly than comments that do not indicate conflict, but noted that readers reported learning more from constructive conflict comments than comments without conflict or without a good resolution. Participants in this study were also aware of the effects comments had on their perceptions of proposal quality, reporting significantly different answers (in the aligned direction) about how reading comments raised and/or lowered their perception of quality, when such differences were observed between their ratings.

I also observed that although the presence of comments may affect perceptions of the proposal the comment is on, that effect **does not** carry over to a second proposal read and judged in quick succession (in this case, on the same Web page), even when the second proposal is by the same author or about a similar mix of topics, and that connection between the proposals is called out with underlined bold header font. Participants in this experiment were apparently able to evaluate the second proposal independently after whatever effects the comment may have had on their perception of the first proposal. This result is encouraging for the future development of platforms that crowdsource proposals about how to solve some particular challenge, where readers may be evaluating those proposals either as part of a platform-hosted contest or for their own reasons (e.g. deciding what proposals to back or adopt in practice).

5. Conclusions

5.1: Brief Summary of Study Findings

As summarized in the introduction above, this thesis contributes to our understanding of human factors in the design of platforms for large-scale online collaboration with peer-generated content. It does so using a range of qualitative and quantitative methods, and makes some methodological contributions (particularly in Study 2) which can be used for future study in this area.

Study 1 identifies particular personas and tasks that are relevant for evaluating whether or not a tool that supports exploratory navigation supports participants' goals, motivations, and helps overcome barriers to those goals. Study 2 suggests that in most cases, similarity links based on an LDA model can have a reasonable level of agreement with human perceptions of similarity, though this algorithm leaves room for improvement. That study also identifies specific areas in which human disagreement with algorithms may be more likely, and contributes a method for testing potential future systems. Study 3 considers how perception of Wikipedia content changes in light of what is presented regarding the discussion behind the content. It finds that perceptions of quality are lower in the presence of any discussion than without (a "sausage effect"), that these effects are stronger if the discussion contains conflict especially if that conflict is not resolved well, but that participants thought they had been affected by the discussion in the opposite direction. Study 4 tests these interesting findings in greater detail, finding that they do not all extend to content accurately framed as a proposal (e.g. the "sausage effect" does not), and that effects that remain in that context (such as the negative effects of conflict) do not extend to a second proposal even when that second proposal is by the same author or about similar topics.

5.2: Discussion of Contributions

While large-scale observational data sets for many online interaction and content production platforms are now available, observational data does not readily permit a distinction between

perceptions based on underlying qualities from those based on other factors that may affect perceptions (Muchnik et al., 2013, p. 647). Controlled experiments with random assignment of the factor being examined, as done here, allows causal conclusions about the differences between levels of the randomly assigned factor.

Especially the latter studies in this thesis contribute to a small set of related studies in the field. For example, Steinfeld, Samuel-Azran, and Lev-On recently published what they described as the first study to examine how readers' perceptions of news articles' quality is affected by a set of comments presented below the article (2016). The study used eye trackers and post-study interviews to measure the attention users paid to the comments. Its setting intended to capture participants' common views about comments sections in online news sites, and found that most participants did not even read the comments, fewer than one in ten read any comments in detail, and even in those cases readers (undergraduates, mostly new first-years, at an Israeli college) often heavily discounted the content of the comments based on stereotypes about people who write them. As a result, that study found that the comments did not influence participants' evaluations of the articles, but drew a primary conclusion emphasizing the need to continue to "map the interplay between user comments and public opinion across various topics and domains" (2016, p. 72).

Past work at *CHI,* for example, has looked at crowdsourced visual presentations of data and how comments on that data by other users affects later users' quantitative perceptions of graphical information. For example, Hullman, Adar, & Shah (2011) asked Turkers to judge proportions or linear association strengths from charts, and found that these judgments were affected by a "social histogram" putatively showing prior viewers' estimates. However, the effects of biased information did not carry over to subsequent chart-perceiving tasks (Hullman et al., 2011, p. 1465). That paper's Future Work section suggests exploring tasks more difficult than perception of quantitative information from the charts used there, where social influence is more likely to be observed. In Study 4, I explore the task of

evaluating the quality of a textually summarized proposal for addressing some aspect of a large, complex problem. My results are consistent with that prior work about social influence on the perception of quantitative content (Hullman et al., 2011), extending that result into more subjective judgments about content quality.

Social influence signals (e.g. comments) affect readers' perception of the content most directly related to the influence signals (e.g. proposals), as evidenced here. These experiments strengthen the call to be aware of these effects and consider them when designing such platforms and the "need to form new theories and models that explain the impact of social processes on community-driven visualization environments and lead to new systems" (Hullman et al., 2011, p. 1469).

5.3: Future Work

5.3.9: Anchor points

Study 4 found results for H1 and H6 that were different from Study 3, despite a strong degree of similarity in task setup and associated details. When considering differences in task setup that may have produced these differences in results, the most likely source seems to be the difference in materials used, i.e. Climate CoLab proposals (and proposal comments) compared to Wikipedia articles (and Talk page discussions).

In contrast to participants in Study 4 based on Climate CoLab proposals, participants in Study 3 based on Wikipedia materials were generally unaware of the effect, particularly the between-subjects primary observation that the presence of even non-conflict comments (in Study 3) led to lower perceptions of article quality. The difference in Study 4 H6 suggests that a different psychological process may have been operating with the different materials, and this may have produced the difference in Study 4 H1.

For example, it may be that in Study 3, the presence of comments caused participants to become more acutely aware of Wikipedia articles' draft-in-progress status and anchor their initial quality measurements based on that category, while participants who saw no discussion evaluated quality beginning from an anchor point more applicable to a reference work perceived as polished or complete. In Study 4, expectation-driven anchor points for the quality of a document described only as a "proposal," even without comments, may have already been lower and perhaps more in line with a draft or work-in-progress than a well-regarded reference work, so the mere presence of comments may not have lowered an initial anchor point. This could explain the finding of no support for Study 4 H1. To determine if this indeed the case, future work would need to explore how anchor points are set for evaluation and how the way people perceive discussion may differ depending on the status of the artifact being discussed.

In a set of experiments, Galak and Nelson (2011) asked readers to evaluate the quality of short stories, experimentally varying the fluency of the text through presentation factors like a more compressed font or asking participants to furrow their brow while reading. They find that the quality rating effects caused by differences in fluency vary as a function of what the reader expects from the reading and the anticipated purpose of the reading (e.g. conveying information contrasted with maximizing enjoyment). It is possible that different purposes for reading cause different expectations, and thus different anchor points from which perceptions are adjusted, between Wikipedia articles and Climate CoLab proposals.

Work in this area relates to the Affective Expectation Model (Wilson, Lisle, Kraft, & Wetzel, 1989), a theory about how people's expectations affects their subjective judgments. In this theory, people who have expectations about the quality value of some content (in the original paper, the humor value of comics) quickly check for features in given material that match their expectations and if found, perceive qualities based on those expectations, even inaccurately. This is a more specific version of schema

theory applied to affect (Wilson et al., 1989, p. 524). According to this theory, people making more specific evaluations (such as with the multidimensional rating scale readers evaluated on here) are more likely to notice discrepancies between their expectations and the actual material being evaluated (Wilson et al., 1989, p. 524), so the most surprising aspects of the theory are unlikely to be responsible for the differences observed here. However, a major part of this theory is about how *quickly* and *easily* people form judgments and how deeply their thoughts are engaged in the material (Wilson et al., 1989, p. 528) and measuring this requires more detailed control over the experimental environment (e.g. lab study) as opposed to Mechanical Turk where variance in e.g. response times can be attributed to a wide variety of other causes. Further research is necessary to better understand the expectation factors that could more clearly link these results into theories that explain the results based on reader expectations.

5.3.10: Larger Comment Sets

Because studies [2,] 3 and 4 used randomized controlled experiment, I have some confidence that the observed effects are *caused* by the experimentally assigned factors, such as the presence and content of the comments. However, in the latter two studies each participant saw only one comment or short discussion posted on each proposal, while in real instances of the intended application environment there may be several comments/discussions posted on each proposal (as also seen in Steinfeld et al., 2016), and readers may extract information from properties of the set, such as total/mean length, number of comments, unique participants, valence mean/variance, or other factors based on the interaction between commenters or between commenters and the content authors, etc. My experiments held those factors relatively constant, and does not tell us if larger sets of comments with various properties would have led to different results. Future work would be needed to determine if more and/or more strongly critical comments have a cumulative effect.

5.3.11: Additional Measures and Modalities

Study 2 compared human perceptions of text similarity to an algorithmic measure based on LDA, which was commonly used, but which treats information stored in word proximity in a coarse binary manner, learning relationships based only on if words are often in the same *document* together or not. New strategies such as word embeddings are now making potentially better use of training information about the meaning of words from context. Crowdsourcing approaches are also being used to generate idea maps facilitating similarity-based navigability among ideas on platforms for large-scale collaborative innovation (Siangliulue, 2017). These techniques could be evaluated using methods that Study 2 contributed. They might also be combined in ways that leverage potentially complementary strengths of crowdsourced and algorithmic approaches.

Text is also a relatively limited modality, and new methods are becoming available for understanding topical and semantic content of images (through advances in computer vision) and even videos (which reportedly make up most internet traffic). For example, a recent (2017) thesis by Lu Jiang describes strategies for efficient, adaptive semantic querying across large (e.g. 100M) video collections. Additional work is being done (e.g. Traina & Traina, 2017) and should be done to enhance navigability between content of different media types, taking advantage of these advances in technology.

Finally, although much work seeks to enhance navigability by similarity comparison, relationships other than similarity may be important for certain key tasks. For example, in a task of attempting to integrate multiple proposals from a large set, it would be very helpful to know which other proposals might both be compatible with and have strengths that help overcome gaps or weaknesses in any selected proposal. Complementarity and intentional diversity can be important to inspiring new contributions as well (Siangliulue, 2015; Siangliulue, Arnold, Gajos, & Dow, 2015). Complementarity measures may require additional advances in natural language processing technology and perhaps greater domain expertise / customization to specific domains, but could be quite useful.

5.4: Broader Implications

Late in the Industrial Revolution, when factory owners replaced centralized mechanical energy with electrical motors, they retained the floor layouts that had been constrained by physical requirements of moving energy to machines from centralized steam pipes or rotating shafts. It wasn't until decades later when managers, paying attention to human factors, discovered the greatest strength this new energy source provided: the flexibility to rearrange what was happening, form a more direct assembly line, and achieve a level of productive performance that was not possible before (Duhigg, 2016).

In the ongoing digital revolution, the Internet and related technologies replace analogous components of how people work together in "traditional" organizations to achieve joint objectives. For example, our collaboration support tools replace paper mail with e-mail. They provide virtual meeting rooms and less formal chat settings. They enable situational awareness and can help support remote collaboration. They automate many routine tasks and improve information-sharing efficiencies in some ways that are analogous to the initial advantages of factory electrification. However, it is only recently (decades after the ARPANET was established) that we are fundamentally reimagining new ways to organize effort (in this case, knowledge work) to take advantage of the tremendous value in diversity and "long tail" distributions of interest, skill, availability, etc., finding valuable gains in accomplishing what was not feasible before. Growing access to diverse human capital is driving changes in social and organizational structures (Towne, 2009), and being able "to cultivate the cross-fertilization of ideas by creating the right kinds of infrastructures and incentives for information exchange" to make the most out of it will be one of the most important managerial skills of the future (Malone, 2004, pp. 157–158).

Open-source software and Wikipedia provide examples of how valuable new forms of working together at large scale can be, while still building largely from concepts like peer review and small-group interaction patterns adapted from more traditional ways of producing conceptually similar outputs. New work in "flash organizations" now takes advantage of technology-enabled flexibility and diversity,

to more quickly solve challenges "traditionally" handled by small to mid-sized teams (Bernstein, 2017). We have yet to harness the now-visible potential of using billions of connected active perspectives to solve the grand challenges that humanity collectively still faces. Some of these challenges that require collaboration at very large scale in order to reach a solution are increasingly urgent.

Platforms intended to support very large scale collaboration are now being built. This work should be done with an understanding of the humans the technology is designed to support, on an individual level and more importantly on a collective level. Human factors affect what people perceive in the material and environments they are working with. These perceptions, and the factors that affect them, can have significant effects on what people trust a process to be able to accomplish, and on real-world actions that impact lives well beyond the small set of people making decisions or designing supports for such decisions. Contextual factors as subtle as the *order* in which supposedly independent judgments are made can distort markets (Hartzmark & Shue, 2016) and affect even expert decision-makers such as major league baseball umpires, loan officers, and asylum immigration judges (Chen, Moskowitz, & Shue, 2016). It is important to understand context effects like these, as explored in the work above. More informed design decisions can more intentionally produce desired consequences. Understanding the subtle effects of design decisions on user experience has led to new design capabilities for common spaces in the physical realm (e.g. Trufelman, 2016) and could be even more powerful for designing digital virtual commons. Future work should continue to explore other ways in which design decisions affect users' perceptions and capacities on platforms for large-scale collaboration.

As before, transforming the way we work together and extending the realm of what we can accomplish requires an understanding of how people are affected by design decisions for the environment in which that work occurs. This thesis advances that understanding by a few small steps. The amount of needed work which remains is still high, but the potential impact from advances in this area is even higher.

Glossary

Afford: The Oxford English Dictionary's definition 3b: "To supply, provide, offer (something sought, or something useful or desirable); to be a source of, to be capable of providing or yielding."

Complex: According to the Oxford English Dictionary, "complex" means "not easy to analyze or understand; complicated or intricate." For this thesis, excluding any discussion associated with (Dabbish, Towne, Diesner, & Herbsleb, 2011), "complex issues" are operationalized as those that no single person or small group can wrap their mind around completely, or fully understand. (Though Engelbart's definition in the framework referred to above is quite different, his final introductory paragraph clearly anticipates "complexity" as defined here (1963, p. 2)). This level of complexity is associated with large systems with many elements and relationships or dependencies between them. It appears that these are the types of issues that could be most fruitfully explored through online deliberation, because (a) existing approaches to decision-making that require decision-makers to fully understand a problem do not adequately address these problems, by definition, and (b) the scale and complexity of online deliberation can potentially grow much larger and faster than the capacities of any individual or group, possibly enough to reach the scale and complexity of the problems themselves. Rittel & Webber (1973) described a "wicked" subset of these problems in greater detail, noting the particular scale and complexity issues that accompany them. A later work described "Ultra-large-scale (ULS)" systems in similar terms (Feiler et al., 2006).

Emergent: Using the definition of Deese, "An emergent phenomenon is something that is not predictable or understandable from the properties of the components taken individually and independently" (Deese, 1972, p. 99).

Ideation: As used here, "ideation" refers to the process or activity of generating and collecting new ideas. **Ideation tools** or **platforms** as referenced here generally focus on the collection (aggregation) of ideas than on facilitating the mental processes that lead to creative thinking and users' generation of

those ideas. Management of and navigability in that set of ideas is then a challenge; see especially the "0.3.3: Ideation Tools" section above for more details.

Policy: For the purposes of this thesis, a "policy decision" is any decision that is intended to have effects that extend well beyond the decision-maker or decision-making group; where the set of direct stakeholders in the outcome of a decision is much larger than the set of people making that decision. For the purposes of this thesis, a legislature's vote to temporarily adjourn is not a policy decision, but its vote on that government's budget is; a supercomputer-consortium committee's decision to change their meeting support technology is not a policy decision but their choice of an authentication technology on the supercomputers is.

Theory: Here, theory is defined broadly to mean a conceptual model on which various findings converge, borrowing from the definition of Mook (1983).

References

Andrzejewski, D., & Zhu, X. (2009). Latent Dirichlet Allocation with Topic-in-set Knowledge. In *Proceedings of the NAACL HLT 2009 Workshop on Semi-Supervised Learning for Natural Language Processing* (pp. 43–48). Stroudsburg, PA, USA: Association for Computational Linguistics. Retrieved from http://dl.acm.org/citation.cfm?id=1621829.1621835

Andrzejewski, D., Zhu, X., Craven, M., & Recht, B. (2011). A Framework for Incorporating General Domain Knowledge into Latent Dirichlet Allocation Using First-order Logic. In *Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence - Volume Two* (pp. 1171–1177). Barcelona, Catalonia, Spain: AAAI Press. https://doi.org/10.5591/978-1-57735-516-8/IJCAI11-200

Appelbaum, Y. (2012, May 15). How the Professor Who Fooled Wikipedia Got Caught by Reddit. *The Atlantic*. Retrieved from http://www.theatlantic.com/technology/archive/2012/05/how-the-professor-who-fooled-wikipedia-got-caught-by-reddit/257134/

Arazy, O., & Nov, O. (2010). Determinants of Wikipedia Quality: The Roles of Global and Local Contribution Inequality. In *Proceedings of the 2010 ACM conference on Computer supported cooperative work* (pp. 233–236). New York, NY, USA: ACM. https://doi.org/10.1145/1718918.1718963

Bailey, B. P., & Horvitz, E. (2010). What's Your Idea?: A Case Study of a Grassroots Innovation Pipeline Within a Large Software Company. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 2065–2074). New York, NY, USA: ACM. https://doi.org/10.1145/1753326.1753641

Bak, R. (2011). *The Big Jump: Lindbergh and the Great Atlantic Air Race*. Hoboken, N.J: John Wiley & Sons. Retrieved from https://books.google.com/books?id=wW56kEOOb1YC

Baldwin, P. (2010). What Is Debategraph? and Debategraph Data Structure. Retrieved from http://debategraph.org/BaldwinArticlesAboutDebategraph

Bank, C., & Cao, J. (2015). *The Guide to Usability Testing*. Mountain View, CA: UXPin. Retrieved from http://www.uxpin.com/guide-to-usability-testing.html

Bates, M. J. (1989). The Design of Browsing and Berrypicking Techniques for the Online Search Interface. *Online Review*, *13*(5), 407–424. https://doi.org/10.1108/eb024320

Berners-Lee, T. (1989). *Information Management: A Proposal* (Proposal No. TBL-900620). CERN. Retrieved from http://cds.cern.ch/record/369245/files/dd-89-001.pdf

Bernstein, M. (2017, February). *In A Flash: Crowdsourcing Organizations, Collaborations, and Research*. Presented at the HCII Seminar Series, Pittsburgh, PA. Retrieved from https://www.hcii.cmu.edu/news/seminar/event/2017/02/flash-crowdsourcing-organizations-collaborations-and-research

Bikhchandani, S., Hirshleifer, D., & Welch, I. (1992). A Theory of Fads, Fashion, Custom, and Cultural Change as Informational Cascades. *Journal of Political Economy*, *100*(5), 992–1026.

Bisson, S., Flaschen, M., Giner, P., Horn, D., Kattouw, R., Mehta, K., … Tonkovidova, E. (2015, May 28). Flow. Retrieved from https://www.mediawiki.org/w/index.php?title=Flow&oldid=1656909

Bjelland, O. M., & Wood, R. C. (2008, October 1). An Inside View of IBM's "Innovation Jam." Retrieved September 6, 2013, from http://sloanreview.mit.edu/article/an-inside-view-of-ibms-innovation-jam/

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. *J. Mach. Learn. Res.*, *3*, 993–1022.

Bonabeau, E. (2009). Decisions 2.0: The Power of Collective Intelligence. *MIT Sloan Management Review*. Retrieved from http://sloanreview.mit.edu/article/decisions-20-the-power-of-collective-intelligence/

Boyack, K. W., Newman, D., Duhon, R. J., Klavans, R., Patek, M., Biberstine, J. R., … Börner, K. (2011). Clustering More than Two Million Biomedical Publications: Comparing the Accuracies of Nine Text-Based Similarity Approaches. *PLoS ONE*, *6*(3), e18029. https://doi.org/10.1371/journal.pone.0018029

Boyd, R. L., & Pennebaker, J. W. (2015). Did Shakespeare Write Double Falsehood? Identifying Individuals by Creating Psychological Signatures With Text Analysis. *Psychological Science*, 0956797614566658. https://doi.org/10.1177/0956797614566658

Boytsov, L. (2016, June). *Thesis Proposal: Off the Beaten Path: Let's Replace Term-Based Retrieval With k-NN Search*. Presented at the Carnegie Mellon University, Pittsburgh, PA. Retrieved from http://searchivarius.org/personal/leonid-boytsovs-publications

Boytsov, L., Novak, D., Malkov, Y., & Nyberg, E. (2016). Off the Beaten Path: Let's Replace Term-Based Retrieval with k-NN Search. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management* (pp. 1099–1108). New York, NY, USA: ACM. https://doi.org/10.1145/2983323.2983815

Bragdon, A., Zeleznik, R., Reiss, S. P., Karumuri, S., Cheung, W., Kaplan, J., … LaViola, J. J., Jr. (2010). Code Bubbles: A Working Set-based Interface for Code Understanding and Maintenance. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 2503–2512). New York, NY, USA: ACM. https://doi.org/10.1145/1753326.1753706

Brandenburger, A. M., & Nalebuff, B. J. (2011). *Co-Opetition*. Crown Publishing Group. Retrieved from https://books.google.com/books?id=sU2e-piQ3tUC

Brokos, G.-I., Malakasiotis, P., & Androutsopoulos, I. (2016). Using Centroids of Word Embeddings and Word Mover's Distance for Biomedical Document Retrieval in Question Answering. In *arXiv:1608.03905 [cs]*. Retrieved from http://arxiv.org/abs/1608.03905

Buckingham Shum, S. (2008). Cohere: Towards Web 2.0 Argumentation. In *Proceeding of the 2008 conference on Computational Models of Argument* (pp. 97–108). Amsterdam, The Netherlands: IOS Press. Retrieved from http://portal.acm.org/citation.cfm?id=1566134.1566144

Bullard, J. (2014). Values and Negotiation in Classification Work. In *Proceedings of the Companion Publication of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing* (pp. 45–48). New York, NY, USA: ACM. https://doi.org/10.1145/2556420.2556820

Bush, V. (1945, July). As We May Think. *The Atlantic*. Retrieved from http://www.theatlantic.com/magazine/archive/1945/07/as-we-may-think/303881/?single_page=true

Candan, K. S., Di Caro, L., & Sapino, M. L. (2012). PhC: Multiresolution Visualization and Exploration of Text Corpora with Parallel Hierarchical Coordinates. *ACM Trans. Intell. Syst. Technol.*, *3*(2), 22:1–22:36. https://doi.org/10.1145/2089094.2089098

Cao, J. (2015, April 27). User Testing, Explained. Retrieved May 2, 2015, from http://thenextweb.com/creativity/2015/04/27/user-testing-explained/

Carey, D. (2015, July 10). Comment 16: Language Requirement. Retrieved April 26, 2016, from http://climatecolab.org/contests/2016/shifting-behavior-for-a-changing-climate/c/proposal/1320702/tab/COMMENTS#_message_1349447

Carpineto, C., Osiński, S., Romano, G., & Weiss, D. (2009). A Survey of Web Clustering Engines. *ACM Comput. Surv.*, *41*(3), 17:1–17:38. https://doi.org/10.1145/1541880.1541884

Cavalier, R. J. (2011). *Approaching Deliberative Democracy: Theory and Practice*. Carnegie Mellon University Press.

Chang, J., & Blei, D. (2009). Relational Topic Models for Document Networks. In *AISTATS* (Vol. 5, pp. 81–88). Clearwater Beach, Florida USA. Retrieved from http://jmlr.csail.mit.edu/proceedings/papers/v5/chang09a/chang09a.pdf

Chang, J., Boyd-Graber, J., Gerrish, S., Wang, C., & Blei, D. M. (2009). Reading Tea Leaves: How Humans Interpret Topic Models. In *Proc. NIPS*. Vancouver, B.C., Canada. Retrieved from https://nips.cc/Conferences/2009/Program/event.php?ID=1812

Chen, D. L., Moskowitz, T. J., & Shue, K. (2016). Decision Making Under the Gambler's Fallacy: Evidence from Asylum Judges, Loan Officers, and Baseball Umpires. *The Quarterly Journal of Economics*, *131*(3), 1181–1242. https://doi.org/10.1093/qje/qjw017

Chesney, T. (2006). An Empirical Examination of Wikipedia's Credibility. *First Monday*, *11*(11–6). Retrieved from http://firstmonday.org/htbin/cgiwrap/bin/ojs/index.php/fm/article/view/1413

Chuang, J., Ramage, D., Manning, C., & Heer, J. (2012). Interpretation and Trust: Designing Model-driven Visualizations for Text Analysis. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 443–452). New York, NY, USA: ACM. https://doi.org/10.1145/2207676.2207738

Cleaver, L. (2013, August). *JAM: Liam Cleaver at TEDxLitchfieldED*. Presented at the TEDxLitchfieldED, Southbury, CT. Retrieved from https://www.youtube.com/watch?v=UCAval4JA8o

Climate CoLab. (2016, March 2). How Will Proposals Be Judged? Contest Rules. Retrieved March 2, 2016, from http://climatecolab.org/web/guest/resources/-/wiki/Main/contest+rules#Howwillproposalsbejudged

Climate CoLab. (n.d.-a). Climate CoLab Advisors. Retrieved September 18, 2014, from http://climatecolab.org/web/guest/advisors/-/wiki/Main/Climate+CoLab+Advisors

Climate CoLab. (n.d.-b). Climate CoLab Fellows. Retrieved September 18, 2014, from http://climatecolab.org/resources/-/wiki/Main/Climate+CoLab+Fellows

Climate CoLab. (n.d.-c). Climate CoLab Judges. Retrieved September 18, 2014, from http://climatecolab.org/resources/-/wiki/Main/Climate+CoLab+Judges

Climate CoLab. (n.d.-d). Community Philosophy and Policies. Retrieved September 19, 2014, from http://climatecolab.org/web/guest/resources/-/wiki/Main/Community%20philosophy%20and%20policies

Climate CoLab. (n.d.-e). Contest Rules. Retrieved September 19, 2014, from http://climatecolab.org/web/guest/resources/-/wiki/Main/2014+contest+rules

Climate CoLab. (n.d.-f). Crowdsourcing. Retrieved September 19, 2014, from http://climatecolab.org/web/guest/crowdsourcing

Climate CoLab. (n.d.-g). Selecting Semi-Finalists. Retrieved September 19, 2014, from http://climatecolab.org/web/guest/resources/-/wiki/Main/Selecting+Semi-Finalists

Cohen, J. (1969). *Statistical Power Analysis for the Behavioral Sciences*. New York: Academic Press.

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences*. New York, NY, USA: L. Erlbaum Associates.

Cohen, J. (1992). A Power Primer. *Psychological Bulletin*, *112*(1), 155–159. https://doi.org/10.1037/0033-2909.112.1.155

Collective Intelligence 2012. (2012). Presented at the Collective Intelligence 2012, Cambridge, MA. Retrieved from http://www.ci2012.org/

Conklin, J. (1987). *A Survey of Hypertext* (Software Technology Program No. MCC TR STP-356-86, Rev. 2) (p. 40). Austin, TX: Microelectronics and Computer Technology Corporation. Retrieved from http://csis.pace.edu/~marchese/CS835/Lec3/conklin95survey.pdf

Conklin, J., & Begeman, M. L. (1988). gIBIS: A Hypertext Tool for Exploratory Policy Discussion. *ACM Transactions on Office Information Systems*, *6*(4), 303–331. https://doi.org/10.1145/58566.59297

Convertino, G. (2013, May). Large-Scale Idea Management and Deliberation Systems Workshop. Retrieved June 2, 2013, from http://comtech13.xrce.xerox.com/comtech13.html

Corbin, J. M., & Strauss, A. (2014). *Basics of Qualitative Research: Techniques and Procedures for Developing Grounded Theory* (4th ed.). Thousand Oaks, CA: SAGE Publications, Inc. Retrieved from http://www.amazon.com/Basics-Qualitative-Research-Techniques-Procedures/dp/1412997461/ref=pd_cp_b_0

Crocker, S. (2016, October 1). Cheers to the Multistakeholder Community. Retrieved November 1, 2016, from https://www.icann.org/news/blog/cheers-to-the-multistakeholder-community

Cui, W., Qu, H., Zhou, H., Zhang, W., & Skiena, S. (2012). Watch the Story Unfold with TextWheel: Visualization of Large-Scale News Streams. *ACM Trans. Intell. Syst. Technol.*, *3*(2), 20:1–20:17. https://doi.org/10.1145/2089094.2089096

Cutting, D. R., Karger, D. R., Pedersen, J. O., & Tukey, J. W. (1992). Scatter/Gather: A Cluster-based Approach to Browsing Large Document Collections. In *Proceedings of the 15th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 318–329). New York, NY, USA: ACM. https://doi.org/10.1145/133160.133214

Dabbish, L., Stuart, C., Tsay, J., & Herbsleb, J. (2012). Social Coding in GitHub: Transparency and Collaboration in an Open Software Repository. In *Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work* (pp. 1277–1286). New York, NY, USA: ACM. https://doi.org/10.1145/2145204.2145396

Dabbish, L., Towne, W. B., Diesner, J., & Herbsleb, J. (2011). Construction of Association Networks from Communication in Teams Working on Complex Projects. *Statistical Analysis and Data Mining*, *4*(5), 547–563. https://doi.org/10.1002/sam.10135

Daniel, A. (2014, November 25). An MIT Project Crowdsources Local Solutions in the Fight Against Climate Change. *Public Radio International's The World*. Cambridge, MA. Retrieved from http://www.pri.org/stories/2014-11-25/mit-project-crowdsources-local-solutions-fight-against-climate-change

DARPA. (2014, March 13). The DARPA Grand Challenge: Ten Years Later. Retrieved April 19, 2015, from http://www.darpa.mil/newsevents/releases/2014/03/13.aspx

Davies, T. (2011, March 7). Online Deliberation Conferences. Retrieved May 13, 2011, from http://online-deliberation.net/

De Liddo, A., & Buckingham Shum, S. (2010). Capturing and Representing Deliberation in Participatory Planning Practices. Presented at the Fourth International Conference on Online Deliberation (OD2010), Leeds, UK. Retrieved from http://ics.leeds.ac.uk/sub1.cfm?pbcrumb=3rd%20February%202010

De Liddo, A., & Buckingham Shum, S. (2013). The Evidence Hub: Harnessing the Collective Intelligence of Communities to Build Evidence-Based Knowledge (p. 8). Presented at the Large Scale Ideation and Deliberation Workshop, Munich, Germany. Retrieved from http://comtech13.xrce.xerox.com/papers/paper3_liddo_%20shum.pdf

de Looper, C. (2016, March 10). Wikipedia's new iOS app will give you the information you're looking for. Retrieved March 11, 2016, from http://www.digitaltrends.com/mobile/new-wikipedia-ios-app/

Dean, J., & Henzinger, M. R. (1999). Finding related pages in the World Wide Web. *Computer Networks*, *31*(11–16), 1467–1479. https://doi.org/10.1016/S1389-1286(99)00022-5

Deese, J. (1972). *Psychology as Science and Art*. Harcourt Brace Jovanovich.

DigitalGov. (2015, April 19). Challenges & Prizes Community. Retrieved April 20, 2015, from http://www.digitalgov.gov/communities/challenges-prizes-community/

Dinakar, K., Jones, B., Lieberman, H., Picard, R., Rose, C., Thoman, M., & Reichart, R. (2012). You Too?! Mixed-Initiative LDA Story Matching to Help Teens in Distress. In *Proceedings of the Sixth International AAAI Conference on Weblogs and Social Media* (pp. 74–81). Dublin, Ireland: AAAI. Retrieved from http://www.aaai.org/ocs/index.php/ICWSM/ICWSM12/paper/view/4604

Donahoe, E. (2016, December). *The Future of Internet Governance in a Post-Transition World*. Presented at the Carnegie Colloquium, Pittsburgh, PA. Retrieved from https://www.youtube.com/watch?v=RH2xfmFVij8

Dorgelo, C. (2014, January 23). Challenge.gov Wins "Innovations in American Government" Award. Retrieved April 19, 2015, from http://www.whitehouse.gov/blog/2014/01/23/challengegov-wins-innovations-american-government-award

Duhaime, E. P., Olson, G. M., & Malone, T. W. (2015a). Broad Participation in Collective Problem Solving Can Influence Participants and Lead to Better Solutions: Evidence from the MIT Climate CoLab. Presented at the Collective Intelligence 2015, Santa Clara, CA. Retrieved from http://sites.lsa.umich.edu/collectiveintelligence/wp-content/uploads/sites/176/2015/02/Duhaime-Olson-and-Malone-CI-2015-Abstract.pdf

Duhaime, E. P., Olson, G. M., & Malone, T. W. (2015b). *Broad Participation in Collective Problem Solving Can Influence Participants and Lead to Better Solutions: Evidence from the MIT Climate CoLab* (Working Paper No. 2015-02). Cambridge, MA: MIT Center for Collective Intelligence. Retrieved from http://cci.mit.edu/working_papers_2012_2013/duhaime%20colab%20wp%206-2015%20final.pdf

Duhigg, C. (2016, April). *How to Be More Productive*. Presented at the Freakonomics. Retrieved from http://freakonomics.com/podcast/how-to-be-more-productive/

Dumais, S. T., Furnas, G. W., Landauer, T. K., Deerwester, S., & Harshman, R. (1988). Using Latent Semantic Analysis to Improve Access to Textual Information. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 281–285). New York, NY, USA: ACM. https://doi.org/10.1145/57167.57214

Dunn, J., Smith, J. M., Towne, W. B., Michael, G., Dagostino, N., Leiva, P., … Park, D. (2014, October). *NextGen@ICANN Presentations*. Presented at the ICANN51, Los Angeles, CA. Retrieved from http://la51.icann.org/en/schedule/thu-nextgen

Easterday, M. W., Kanarek, J. S., & Harrell, M. (2009). Design Requirements of Argument Mapping Software for Teaching Deliberation. In T. Davies & S. P. Gangadhara (Eds.), *Online Deliberation: Design, Research, and Practice.* (pp. 317–323). CSLI Publications.

Eckert, J. B. (1995). *The Super Efficient Refrigerator Program: Case Study of a Golden Carrot Program* (UC Category 1600 No. NREL/TP-461-7281 DE95009255). Golden, Colorado: National Renewable Energy Laboratory. Retrieved from http://www.nrel.gov/docs/legosti/old/7281.pdf

Egan, D. E., Remde, J. R., Landauer, T. K., Lochbaum, C. C., & Gomez, L. M. (1989). Behavioral Evaluation and Analysis of a Hypertext Browser. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 205–210). New York, NY, USA: ACM. https://doi.org/10.1145/67449.67490

El-Yaniv, R., Fine, S., & Tishby, N. (1997). Agnostic Classification of Markovian Sequences. In M. I. Jordan, M. J. Kearns, & S. A. Solla (Eds.), *Advances in Neural Information Processing Systems 10* (pp. 465–471). MIT Press. Retrieved from http://papers.nips.cc/paper/1376-agnostic-classification-of-markovian-sequences.pdf

Endres, D. M., & Schindelin, J. E. (2003). A new metric for probability distributions. *IEEE Transactions on Information Theory*, *49*(7), 1858–1860. https://doi.org/10.1109/TIT.2003.813506

Engelbart, D. C. (1963). A Conceptual Framework for the Augmentation of Man's Intellect. In P. W. Howerton & D. C. Weeks (Eds.), *Vistas in Information Handling* (Vol. 1, pp. 1–29). Stanford Research Institute, Menlo Park, California: Spartan Books. Retrieved from http://www.amazon.co.uk/Information-Handling-General-Howerton-Associate/dp/B000XJ3V9U

Farrahi, K., & Gatica-Perez, D. (2011). Discovering Routines from Large-scale Human Locations Using Probabilistic Topic Models. *ACM Trans. Intell. Syst. Technol.*, *2*(1), 3:1–3:27. https://doi.org/10.1145/1889681.1889684

Faruqui, M., Dodge, J., Jauhar, S. K., Dyer, C., Hovy, E., & Smith, N. A. (2015). Retrofitting Word Vectors to Semantic Lexicons. In *Proceedings of NAACL*. Retrieved from http://aclweb.org/anthology/N/N15/N15-1184.pdf

Faul, F., Erdfelder, E., Buchner, A., & Lang, A.-G. (2009). Statistical Power Analyses Using G*Power 3.1: Tests for Correlation and Regression Analyses. *Behavior Research Methods*, *41*(4), 1149–1160. https://doi.org/10.3758/BRM.41.4.1149

Feiler, P., Gabriel, R. P., Goodenough, J., Linger, R., Longstaff, T., Kazman, R., … Wallnau, K. (2006). *Ultra-Large-Scale Systems: The Software Challenge of the Future*. (B. Pollak, Ed.). Pittsburgh, PA: Software Engineering Institute, Carnegie Mellon University. Retrieved from http://www.amazon.com/Ultra-Large-Scale-Systems-Software-Challenge-Future/dp/0978695607/

Fleet, D. D. V., & Atwater, L. (1997). Gender Neutral Names: Don't Be So Sure! *Sex Roles*, *37*(1–2), 111–123. https://doi.org/10.1023/A:1025696905342

Fogg, B. J., Soohoo, C., Danielson, D. R., Marable, L., Stanford, J., & Tauber, E. R. (2003). How Do Users Evaluate the Credibility of Web Sites? In *Proceedings of the 2003 Conference on Designing for User Experiences - DUX '03* (p. 1). San Francisco, California. https://doi.org/10.1145/997078.997097

Forman, G., Eshghi, K., & Chiocchetti, S. (2005). Finding Similar Files in Large Document Repositories. In *Proceedings of the Eleventh ACM SIGKDD International Conference on Knowledge Discovery in Data Mining* (pp. 394–400). New York, NY, USA: ACM. https://doi.org/10.1145/1081870.1081916

Fritsch, J., Pesenti, J., Hutchison, K., Cancelliere, R., & Srini, J. (2013, December). *MIT Enterprise Forum: Big Data in Life Sciences*. Panel presented at the MIT Enterprise Forum Pittsburgh, Thermo Fisher Scientific Offices, Pittsburgh PA.

Furnas, G. W., Landauer, T. K., Gomez, L. M., & Dumais, S. T. (1987). The Vocabulary Problem in Human-system Communication. *Commun. ACM*, *30*(11), 964–971. https://doi.org/10.1145/32206.32212

Gaikwad, S., Morina, D., Nistala, R., Agarwal, M., Cossette, A., Bhanu, R., … Bernstein, M. (2015). Daemo: a Self-Governed Crowdsourcing Marketplace. In *Proceedings of the 28th Annual ACM Symposium on User Interface Software & Technology*. Charlotte, NC: ACM. Retrieved from https://uist.acm.org/uist2015/schedule#uistp142

Galak, J., & Nelson, L. D. (2011). The Virtues of Opaque Prose: How Lay Beliefs About Fluency Influence Perceptions of Quality. *Journal of Experimental Social Psychology*, *47*(1), 250–253. https://doi.org/10.1016/j.jesp.2010.08.002

Galton, F. (1907). Vox Populi (The Wisdom of Crowds). *Nature*, *75*(1949), 450–451. https://doi.org/10.1038/075450a0

Garber, M. (2012, May 9). Abraham Lincoln Did Not Invent Facebook: How a Guy and His Blog Fooled the Whole Wide Internet. *The Atlantic*. Retrieved from http://www.theatlantic.com/technology/archive/2012/05/abraham-lincoln-did-not-invent-facebook-how-a-guy-and-his-blog-fooled-the-whole-wide-internet/256945/

Glass, I. (2016, October 21). Seriously? *This American Life*. NPR. Retrieved from http://www.thisamericanlife.org/radio-archives/episode/599/seriously

Goggins, S., Mascaro, C., & Mascaro, S. (2012). Relief Work After the 2010 Haiti Earthquake: Leadership in an Online Resource Coordination Network. In *CSCW* (pp. 57–66). New York, NY, USA: ACM. https://doi.org/10.1145/2145204.2145218

Gretarsson, B., O'Donovan, J., Bostandjiev, S., Höllerer, T., Asuncion, A., Newman, D., & Smyth, P. (2012). TopicNets: Visual Analysis of Large Text Corpora with Topic Modeling. *ACM Trans. Intell. Syst. Technol.*, *3*(2), 23:1–23:26. https://doi.org/10.1145/2089094.2089099

Gustetic, J. (2015, April 17). 21st-Century Public Servants: Using Prizes and Challenges to Spur Innovation. Retrieved April 19, 2015, from http://www.whitehouse.gov/blog/2015/04/17/21st-century-public-servants-using-prizes-and-challenges-spur-innovation

Harari, Y. (2016, March). *Why Did Humans Become The Most Successful Species On Earth?* Presented at the TEDGlobalLondon, London. Retrieved from http://www.npr.org/2016/03/04/468882620/why-did-humans-become-the-most-successful-species-on-earth

Harrell, M. (2005a). Using Argument Diagramming Software in the Classroom. *Teaching Philosophy*, *28*(2), 163–177. https://doi.org/10.5840/teachphil200528222

Harrell, M. (2005b). Using Argument Diagrams to Improve Critical Thinking Skills in 80-100 What Philosophy Is. *Department of Philosophy*. Retrieved from http://repository.cmu.edu/philosophy/120

Harrell, M. (2011). Argument Diagramming and Critical Thinking in Introductory Philosophy. *Higher Education Research & Development*, *30*(3), 371–385. https://doi.org/10.1080/07294360.2010.502559

Hartzmark, S. M., & Shue, K. (2016). *A Tough Act to Follow: Contrast Effects in Financial Markets* (SSRN Scholarly Paper No. ID 2613702). Rochester, NY: Social Science Research Network. Retrieved from https://papers.ssrn.com/abstract=2613702

Haveliwala, T. H., Gionis, A., Klein, D., & Indyk, P. (2002). Evaluating Strategies for Similarity Search on the Web. In *Proceedings of the 11th International Conference on World Wide Web* (pp. 432–442). New York, NY, USA: ACM. https://doi.org/10.1145/511446.511502

He, H., He, P., Gao, J., & Huang, C. (2002). Performance of Two Information Retrieval Systems in Chinese IR: SMART System and Okapi System. In *TENCON '02. Proceedings. 2002 IEEE Region 10 Conference on Computers, Communications, Control and Power Engineering* (Vol. 1, pp. 465–468 vol.1). https://doi.org/10.1109/TENCON.2002.1181314

Hearst, M. A., & Pedersen, J. O. (1996). Reexamining the Cluster Hypothesis: Scatter/Gather on Retrieval Results. In *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 76–84). New York, NY, USA: ACM. https://doi.org/10.1145/243199.243216

Hirshleifer, D., & Hong Teoh, S. (2003). Herd Behaviour and Cascading in Capital Markets: a Review and Synthesis. *European Financial Management*, *9*(1), 25–66. https://doi.org/10.1111/1468-036X.00207

Hopkins, M. S., & Malone, T. W. (2009). All Together Now (or, Can Collective Intelligence Save the Planet?). *MIT Sloan Management Review*. Retrieved from http://sloanreview.mit.edu/article/can-collective-intelligence-save-the-planet/

Howard, D. J., & Kerin, R. A. (2011). The Effects of Name Similarity on Message Processing and Persuasion. *Journal of Experimental Social Psychology*, *47*(1), 63–71. https://doi.org/10.1016/j.jesp.2010.08.008

Howe, J. (2008). *Crowdsourcing: Why the Power of the Crowd Is Driving the Future of Business*. Crown Business.

Huang, Y., & Mitchell, T. (2007). A Framework for Mixed-Initiative Clustering. In *North East Student Colloquium on Artificial Intelligence (NESCAI 2007)*. Ithaca, NY. Retrieved from http://www.cs.cornell.edu/Conferences/nescai/nescai07/

Hudson, W. (2014). Card Sorting. In M. Soegaard & R. F. Dam (Eds.), *The Encyclopedia of Human-Computer Interaction* (2nd ed.). Aarhus, Denmark: The Interaction Design Foundation. Retrieved from https://www.interaction-design.org/encyclopedia/card_sorting.html

Hullman, J., Adar, E., & Shah, P. (2011). The Impact of Social Information on Visual Judgments. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 1461–1470). New York, NY, USA: ACM. https://doi.org/10.1145/1978942.1979157

Hwang, K. O., Ottenbacher, A. J., Green, A. P., Cannon-Diehl, M. R., Richardson, O., Bernstam, E. V., & Thomas, E. J. (2010). Social Support in an Internet Weight Loss Community. *International Journal of Medical Informatics*, *79*(1), 5–13. https://doi.org/10.1016/j.ijmedinf.2009.10.003

Iandoli, L., Klein, M., & 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*(1), 69–91.

IBM. (2012, March 7). A Global Innovation Jam [CTB14]. Retrieved September 6, 2013, from http://www-03.ibm.com/ibm/history/ibm100/us/en/icons/innovationjam/

IdeaScale. (2012). Save Award 2012. Retrieved June 13, 2013, from http://saveaward2012.ideascale.com/

IdeaScale. (2013, April 8). The Truth About IdeaScale. Retrieved June 2, 2013, from http://dev.ideascale.com/infocomics/

Introne, J. E., & Drescher, M. (2013). Analyzing the flow of knowledge in computer mediated teams. In *Proceedings of the 2013 conference on Computer supported cooperative work* (pp. 341–356). New York, NY, USA: ACM. https://doi.org/10.1145/2441776.2441816

IPCC. (2014). *Impacts, Adaptation and Vulnerability* (Assessment Report No. 5). Intergovernmental Panel on Climate Change. Retrieved from http://www.ipcc.ch/report/ar5/wg2/

Jehn, K. A. (1997). A Qualitative Analysis of Conflict Types and Dimensions in Organizational Groups. *Administrative Science Quarterly*, *42*(3), 530–557. https://doi.org/10.2307/2393737

Jiang, L. (2017, April). *Web-scale Multimedia Search for Internet Video Content*. Carnegie Mellon University, Pittsburgh, PA. Retrieved from https://www.cs.cmu.edu/calendar/thu-2017-04-27-0900/language-technologies-thesis-defense

Johnson, S. (2010). *Where Good Ideas Come From*. Penguin.

Joseph, K., Carley, K. M., & Hong, J. I. (2014). Check-ins in "Blau Space:" Applying Blau's Macrosociological Theory to Foursquare Check-ins from New York City. *ACM Trans. Intell. Syst. Technol.*, *5*(3), 46:1–46:22. https://doi.org/10.1145/2566617

Joshi, P., Parrish, M., & Bauman, A. S. (1993, June 30). A Cool $30 Million: Whirlpool Wins Prize for Designing Environmentally Safe Refrigerator. *Los Angeles Times*. Retrieved from http://articles.latimes.com/1993-06-30/business/fi-8558_1_energy-efficient-refrigerator

Kahneman, D. (2011). *Thinking, Fast and Slow* (1st ed.). Farrar, Straus and Giroux.

Kasof, J. (1993). Sex Bias in the Naming of Stimulus Persons. *Psychological Bulletin*, *113*(1), 140–163. https://doi.org/10.1037/0033-2909.113.1.140

Kaufman, Ben. (2015, February). *Archives & Similars*. Presented at the Quirky Town Meeting, New York, NY, USA. Retrieved from http://www.ustream.tv/recorded/58936980/highlight/603163

Kittur, A., Chi, E. H., & Suh, B. (2008). Crowdsourcing User Studies with Mechanical Turk. In *Proceedings of the Twenty-Sixth Annual Sigchi Conference on Human Factors in Computing Systems* (pp. 453–456). New York, NY, USA: ACM. https://doi.org/10.1145/1357054.1357127

Kittur, A., & Kraut, R. E. (2008). Harnessing the Wisdom of Crowds in Wikipedia: Quality Through Coordination. In *Proceedings of the 2008 ACM conference on Computer supported cooperative work* (pp. 37–46). New York, NY, USA: ACM. https://doi.org/10.1145/1460563.1460572

Kittur, A., Smus, B., Khamkar, S., & Kraut, R. E. (2011). CrowdForge: Crowdsourcing Complex Work. In *Proceedings of the 24th Annual ACM Symposium on User Interface Software and Technology* (pp. 43–52). New York, NY, USA: ACM. https://doi.org/10.1145/2047196.2047202

Kittur, A., Suh, B., & Chi, E. H. (2008). Can You Ever Trust a Wiki? In *Proceedings of the ACM 2008 Conference on Computer Supported Cooperative Work* (p. 477). San Diego, CA, USA. https://doi.org/10.1145/1460563.1460639

Kittur, A., Suh, B., Pendleton, B. A., & Chi, E. H. (2007). He Says, She Says: Conflict and Coordination in Wikipedia. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 453–462). San Jose, California, USA. https://doi.org/10.1145/1240624.1240698

Klein, M. (2009, January 6). Open For Questions – A Critique of Idea Sharing Sites. Retrieved December 4, 2010, from http://markklein.wordpress.com/2009/01/06/a-critique-of-idea-sharing/

Klein, M. (2011a). *How to Harvest Collective Wisdom on Complex Problems: An Introduction to the MIT Deliberatorium*. CCI Working Paper, MIT Center for Collective Intelligence. Retrieved from http://cci.mit.edu/klein/papers/deliberatorium-intro.pdf

Klein, M. (2011b). The MIT deliberatorium: Enabling large-scale deliberation about complex systemic problems. In *2011 International Conference on Collaboration Technologies and Systems (CTS)* (pp. 161–161). https://doi.org/10.1109/CTS.2011.5928678

Klein, M., & Lu, S. C.-Y. (1989). Conflict Resolution in Cooperative Design. *Artificial Intelligence in Engineering*, *4*(4), 168–180. https://doi.org/10.1016/0954-1810(89)90013-7

Ko, A. J., Myers, B. A., Coblenz, M. J., & Aung, H. H. (2006). An Exploratory Study of How Developers Seek, Relate, and Collect Relevant Information during Software Maintenance Tasks. *IEEE Transactions on Software Engineering*, *32*(12), 971–987. https://doi.org/10.1109/TSE.2006.116

Koshman, S., Spink, A., & Jansen, B. J. (2006). Web Searching on the Vivisimo Search Engine. *Journal of the American Society for Information Science and Technology*, *57*(14), 1875–1887. https://doi.org/10.1002/asi.20408

Kraut, R. E., & Resnick, P. (2012). Encouraging Contributions to Online Communities. In *Building Successful Online Communities: Evidence-Based Social Design*. Cambridge, MA: MIT Press. Retrieved from http://kraut.hciresearch.org/content/books

Kriplean, T., Toomim, M., Morgan, J., Borning, A., & Ko, A. (2012). Is This What You Meant?: Promoting Listening on the Web with Reflect. In *Proceedings of the 2012 ACM Annual Conference on Human Factors in Computing Systems* (pp. 1559–1568). New York, NY, USA: ACM. https://doi.org/10.1145/2207676.2208621

Kunz, W., & Rittel, H. W. J. (1970). *Issues As Elements of Information Systems* (Working Paper No. 131). University of California at Berkeley: Institute of Urban and Regional Development.

Lange, J. D. (2014, October). *Newcomer Welcome Session*. Presented at the ICANN51, Los Angeles, CA. Retrieved from http://la51.icann.org/en/schedule/sun-newcomer

Larson, S. (2014, May 13). Climate Colab Thinks You Could Be the One to Fix Global Warming. Retrieved September 5, 2014, from http://grist.org/climate-energy/climate-colab-thinks-you-could-be-the-one-to-fix-global-warming/

Leber, J. (2014, October 27). What Book Should You Read Next? Putting Librarians And Algorithms To The Test. *Fast Company Co.Exist*. Retrieved from http://www.fastcoexist.com/3037181/what-book-should-you-read-next-putting-librarians-and-algorithms-to-the-test

Lee, J. (1990). SIBYL: A Tool for Managing Group Design Rationale. In *Proceedings of the 1990 ACM Conference on Computer-supported Cooperative Work* (pp. 79–92). New York, NY, USA: ACM. https://doi.org/10.1145/99332.99344

Lee, M. D., Pincombe, B. M., & Welsh, M. B. (2005). An Empirical Evaluation of Models of Text Document Similarity. Presented at the XXVII Annual Conference of the Cognitive Science Society, Stresa, Italy: Cognitive Science Society. Retrieved from http://digital.library.adelaide.edu.au/dspace/handle/2440/28910

Leonhardt, D. (2007, January 31). You Want Innovation? Offer a Prize. *The New York Times*. Retrieved from http://www.nytimes.com/2007/01/31/business/31leonhardt.html

Lewis, R., Dunbar, B., & Crusan, J. (2013, November 21). NASA Center of Excellence for Collaborative Innovation [Text]. Retrieved April 20, 2015, from /offices/COECI/index.html

Liljeqvist, B. (2013, October). Spontaneous Order and Mensa. *Mensa Bulletin*, (569), 56–58.

Lin, J. (1991). Divergence Measures Based on the Shannon Entropy. *IEEE Transactions on Information Theory*, *37*(1), 145–151. https://doi.org/10.1109/18.61115

Lindsay, B. (2009). Creating "The Wikipedia of Pros and Cons." In *Proceedings of WikiSym '09*. Orlando, Florida: ACM. https://doi.org/10.1145/1641309.1641358

Liu, J., & Ram, S. (2011). Who Does What: Collaboration Patterns in the Wikipedia and Their Impact on Article Quality. *ACM Trans. Manage. Inf. Syst.*, *2*(2), 11:1–11:23. https://doi.org/10.1145/1985347.1985352

Liu, S., Zhou, M. X., Pan, S., Song, Y., Qian, W., Cai, W., & Lian, X. (2012). TIARA: Interactive, Topic-Based Visual Text Summarization and Analysis. *ACM Trans. Intell. Syst. Technol.*, *3*(2), 25:1–25:28. https://doi.org/10.1145/2089094.2089101

Liu, Z., Zhang, Y., Chang, E. Y., & Sun, M. (2011). PLDA+: Parallel Latent Dirichlet Allocation with Data Placement and Pipeline Processing. *ACM Trans. Intell. Syst. Technol.*, *2*(3), 26:1–26:18. https://doi.org/10.1145/1961189.1961198

Lucassen, T., & Schraagen, J. M. (2010). Trust in Wikipedia: How Users Trust Information From an Unknown Source. In *Proceedings of the 4th Workshop on Information Credibility* (pp. 19–26). https://doi.org/10.1145/1772938.1772944

Mai, J.-E. (2010). Classification in a Social World: Bias and Trust. *Journal of Documentation*, *66*(5), 627–642. https://doi.org/10.1108/00220411011066763

Malone, T. W. (2004). *The Future of Work: How the New Order of Business Will Shape Your Organization, Your Management Style, and Your Life*. Harvard Business School Press.

Malone, T. W., Laubacher, R., & Dellarocas, C. (2010). The Collective Intelligence Genome. *MIT Sloan Management Review*, *51*(3), 21–31.

Malone, T. W., Laubacher, R., & Fisher, L. (2014, January 15). How Millions of People Can Help Solve Climate Change. *PBS NOVA Next*. Retrieved from http://www.pbs.org/wgbh/nova/next/earth/crowdsourcing-climate-change-solutions/

Malone, T. W., Nickerson, J., Laubacher, R., Fisher, L., de Boer, P., Han, Y., & Towne, W. B. (2017). Putting the Pieces Back Together Again: Contest Webs for Large-Scale Problem Solving. In *Proceedings of the ACM 2017 Conference on Computer Supported Cooperative Work* (p. In Press). Portland, Oregon. Retrieved from http://ssrn.com/abstract=2912951

Malouf, R., & Mullen, T. (2007). Graph-Based User Classification for Informal Online Political Discourse. In *Proceedings of the 1st Workshop on Information Credibility on the Web*. Retrieved from http://www-rohan.sdsu.edu/~gawron/mt_plus/readings/sim_readings/malouf-mullen-wicow07.pdf

Mani, I. (2001). Evaluation. In *Automatic Summarization* (Vol. 3, pp. 221–259). Philadelphia: John Benjamins Publishing. Retrieved from http://books.google.com/books?id=WVUfl1JsKVQC&source=gbs_ViewAPI

Mathews, D. (2014). *The Ecology of Democracy* (First edition). Dayton, OH: Kettering Foundation Press.

Mercier, H., & Sperber, D. (2011). Argumentation: Its Adaptiveness and Efficacy. *Behavioral and Brain Sciences*, *34*(02), 94–111. https://doi.org/10.1017/S0140525X10003031

Metzler, D., Dumais, S., & Meek, C. (2007). Similarity Measures for Short Segments of Text. In G. Amati, C. Carpineto, & G. Romano (Eds.), *Advances in Information Retrieval* (pp. 16–27). Springer Berlin Heidelberg. Retrieved from http://link.springer.com/chapter/10.1007/978-3-540-71496-5_5

Michelucci, P., & Dickinson, J. L. (2016). The Power of Crowds. *Science*, *351*(6268), 32–33. https://doi.org/10.1126/science.aad6499

Miniter, K., Kriplean, T., & Toomim, M. (2014). Considerit. Presented at the National Conference on Dialogue & Deliberation, Reston, VA. Retrieved from https://consider.it/

Minor, J. (2016, March 10). New Wikipedia app for iOS puts the joy of exploration in your hand. Retrieved March 11, 2016, from http://blog.wikimedia.org/2016/03/10/new-wikipedia-app-ios/

MIT Sloan Newsroom. (2014). *Collective Intelligence 2014*. Cambridge, MA: MIT Sloan School of Management. Retrieved from http://mitsloan.mit.edu/newsroom/2014-collective-intelligence-conference.php

Mohler, M., & Mihalcea, R. (2009). Text-to-text Semantic Similarity for Automatic Short Answer Grading. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics* (pp. 567–575). Stroudsburg, PA, USA: Association for Computational Linguistics. Retrieved from http://dl.acm.org/citation.cfm?id=1609067.1609130

Mondak, J. J. (1990). Perceived Legitimacy of Supreme Court Decisions: Three Functions of Source Credibility. *Political Behavior*, *12*(4), 363–384.

Montoya-Weiss, M. M., Massey, A. P., & Song, M. (2001). Getting It Together: Temporal Coordination and Conflict Management in Global Virtual Teams. *The Academy of Management Journal*, *44*(6), 1251–1262. https://doi.org/10.2307/3069399

Mook, D. G. (1983). In Defense of External Invalidity. *American Psychologist*, *38*(4), 379–387. https://doi.org/10.1037/0003-066X.38.4.379

Muchnik, L., Aral, S., & Taylor, S. J. (2013). Social Influence Bias: A Randomized Experiment. *Science*, *341*(6146), 647–651. https://doi.org/10.1126/science.1240466

Muller, M., & Chua, S. (2012). Brainstorming for Japan: Rapid Distributed Global Collaboration for Disaster Response. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 2727–2730). New York, NY, USA: ACM. https://doi.org/10.1145/2207676.2208668

Nahm, U. Y. (2004, August). *Text Mining with Information Extraction*. University of Texas at Austin, Austin, TX. Retrieved from http://www.cs.utexas.edu/users/ai-lab/?nahm:phd04

National Telecommunications & Information Administration. (2014). *NTIA Announces Intent to Transition Key Internet Domain Name Functions*. Washington, DC: US Department of Commerce. Retrieved from http://www.ntia.doc.gov/press-release/2014/ntia-announces-intent-transition-key-internet-domain-name-functions

Newman, D., Lau, J. H., Grieser, K., & Baldwin, T. (2010). Automatic evaluation of topic coherence. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics* (pp. 100–108). Stroudsburg, PA, USA: Association for Computational Linguistics. Retrieved from http://dl.acm.org/citation.cfm?id=1857999.1858011

Newton, R. R., & Rudestam, K. E. (1999). *Your Statistical Consultant: Answers to Your Data Analysis Questions*. Thousand Oaks, CA: Sage. Retrieved from https://books.google.com/books?id=Ys8fFM1Ov3IC

Njaa, C. “Chip.” (2015, May 8). Search for Prior Similar (or exactly the same) Submissions. Retrieved May 19, 2015, from http://community.quirky.com/t/search-for-prior-similar-or-exactly-the-same-submissions/4349

Obama Administration. (2015, April 19). Open Government Initiative: Collaboration. Retrieved April 19, 2015, from https://www.whitehouse.gov/node/293771

Optimal Workshop. (2015, May 4). The World leader in Card Sorting Tools. Retrieved May 4, 2015, from https://www.optimalworkshop.com/optimalsort

Ozturk, P., Han, Y., Towne, W. B., & Nickerson, J. V. (2016). Topic Prevalence and Reuse in an Open Innovation Community. In *Collective Intelligence*. New York, NY. Retrieved from https://sites.google.com/a/stern.nyu.edu/collective-intelligence-conference/

Page, S. E. (2008). *The Difference: How the Power of Diversity Creates Better Groups, Firms, Schools, and Societies*. Princeton University Press.

Palen, L., Anderson, K. M., Mark, G., Martin, J., Sicker, D., Palmer, M., & Grunwald, D. (2010). A Vision for Technology-Mediated Support for Public Participation & Assistance in Mass Emergencies & Disasters. In *Proceedings of the 2010 ACM-BCS Visions of Computer Science Conference* (p. 8:1–8:12). Swinton, UK, UK: British Computer Society. Retrieved from http://dl.acm.org/citation.cfm?id=1811182.1811194

Paul, M., & Girju, R. (2009). Cross-Cultural Analysis of Blogs and Forums with Mixed-Collection Topic Models. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 3* (pp. 1408–1417). Stroudsburg, PA, USA: Association for Computational Linguistics. Retrieved from http://dl.acm.org/citation.cfm?id=1699648.1699687

Paul, M. J. (2001). Interactive Disaster Communication on the Internet: A Content Analysis of Sixty-Four Disaster Relief Home Pages. *Journalism & Mass Communication Quarterly*, *78*(4), 739–753. https://doi.org/10.1177/107769900107800408

Penn, C. (1993, April). Super-Efficient Refrigerator Finalists. *Home Energy Magazine*. Retrieved from http://www.homeenergy.org/show/article/id/931

Perez, S. (2016, March 10). Wikipedia's new iOS app focuses on discovery, personalization. Retrieved March 11, 2016, from http://social.techcrunch.com/2016/03/10/wikipedias-new-ios-app-focuses-on-discovery-personalization/

Phillips, K. W. (2014, October 1). How Diversity Makes Us Smarter. Retrieved March 5, 2017, from https://www.scientificamerican.com/article/how-diversity-makes-us-smarter/

Pincombe, B. (2004). *Comparison of Human and Latent Semantic Analysis (LSA) Judgements of Pairwise Document Similarities for a News Corpus* (Intelligence, Surveillance and Reconnaissance Division, Information Sciences Laboratory No. DSTO-RR-0278). Edinburgh, South Australia: Australian Government Department of Defence: Defence Science and Technology Organisation. Retrieved from http://www.dtic.mil/cgi-bin/GetTRDoc?AD=ADA427585

Pirolli, P., Schank, P., Hearst, M., & Diehl, C. (1996). Scatter/Gather Browsing Communicates the Topic Structure of a Very Large Text Collection. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 213–220). New York, NY, USA: ACM. https://doi.org/10.1145/238386.238489

Pirolli, P., Wollny, E., & Suh, B. (2009). So You Know You're Getting the Best Possible Information: A Tool that Increases Wikipedia Credibility. In *Proceedings of the 27th International Conference on Human Factors in Computing Systems* (p. 1505). Boston, MA, USA: ACM. https://doi.org/10.1145/1518701.1518929

Pitsos, E. (2016). *Kialo*. Munich, Germany. Retrieved from https://www.youtube.com/watch?v=MifNyU49_JA&feature=youtu.be

Pomerantz, W. (2006, October). *Advancements Through Prizes*. Presented at the NIAC Annual Meeting, Tuscon, AZ. Retrieved from http://www.niac.usra.edu/files/library/meetings/annual/oct06/Pomerantz_William.pdf

Pritzker, P., Crocker, S., & Chehade, F. (2014, October). *Welcome Ceremony and President's Opening Session*. Presented at the ICANN51, Los Angeles, CA. Retrieved from http://la51.icann.org/en/schedule/mon-welcome

Privitera, G. J. (2015). *Student Study Guide With IBM® SPSS® Workbook for Essential Statistics for the Behavioral Sciences*. SAGE Publications. Retrieved from https://books.google.com/books?id=-kd9BgAAQBAJ

Remde, J. R., Gomez, L. M., & Landauer, T. K. (1987). SuperBook: An Automatic Tool for Information Exploration—Hypertext? In *Proceedings of the ACM Conference on Hypertext* (pp. 175–188). New York, NY, USA: ACM. https://doi.org/10.1145/317426.317440

Richtel, M. (2013, November 9). A Founder of Twitter Goes Long. *The New York Times*. Retrieved from http://www.nytimes.com/2013/11/10/business/a-founder-of-twitter-goes-long.html

Rittel, H. W. J., & Webber, M. M. (1973). Dilemmas in a General Theory of Planning. *Policy Sciences*, *4*, 155–169.

Rose, J., & Sæbø, Ø. (2010). Designing Deliberation Systems. *The Information Society*, *26*(3), 228–240. https://doi.org/10.1080/01972241003712298

Rossiter, J. (2014, September 26). Progress Report: Continued Product Focus. Retrieved January 1, 2015, from http://yahoo.tumblr.com/post/98474044364/progress-report-continued-product-focus

Safire, W. (2009, February 8). On Language: Fat Tail and Crowdsourcing. *The New York Times*, p. MM24. New York, NY, USA.

Salganik, M. J., Dodds, P. S., & Watts, D. J. (2006). Experimental Study of Inequality and Unpredictability in an Artificial Cultural Market. *Science*, *311*(5762), 854.

Salminen, J. (2012). Collective Intelligence in Humans: A Literature Review. In *arXiv:1204.3401 [cs]*. Cambridge, MA. Retrieved from http://arxiv.org/abs/1204.3401

Santillo, N. (2013, September 23). Because Has Been Busy! Retrieved November 1, 2013, from http://blog.teambecause.com/post/62067976813/because-has-been-busy

Santillo, T. J., & Santillo, N. (2013, August 21). Team Because: Alpha Site. Retrieved January 11, 2014, from http://alpha.teambecause.com/#/post/63

Santoso, S. (2015, April). *White House Office of Science and Technology Policy Senior Policy Advisor*. Interview/Group Presentation, Washington, DC.

Schulman, M. (2013, November 22). Founder of Reddit and the Internet's Own Cheerleader. *The New York Times*. Retrieved from http://www.nytimes.com/2013/11/24/fashion/The-Founder-of-Reddit-Alexis-Ohanian-is-The-Internets-Own-Cheerleader.html

Sciullo, M. (2013, April 25). Pew Study: Political Engagement Online Leaps. *Pittsburgh Post-Gazette*, p. A-1. Pittsburgh, PA.

Shalizi, C. R. (n.d.). Links. Retrieved November 23, 2014, from http://vserver1.cscs.lsa.umich.edu/~crshalizi/links.html

Shirky, C. (2005). *Ontology is Overrated -- Categories, Links, and Tags*. O'Reilly ETech Conference & IMCExpo. Retrieved from http://www.shirky.com/writings/ontology_overrated.html

Shirky, C. (2008). *Here Comes Everybody: the Power of Organizing Without Organizations*. Penguin.

Shirky, C. (2010). *Cognitive Surplus: Creativity and Generosity in a Connected Age*. Penguin. Retrieved from https://books.google.com/books?id=m2rRjwEACAAJ

Shlain, T. (2011). *Connected: An Autoblogography About Love, Death & Technology*. Moxie Institute. Retrieved from http://connectedthefilm.com/

Shneiderman, B. (2000). Designing Trust into Online Experiences. *Communications of the ACM*, *43*(12), 57–59. https://doi.org/10.1145/355112.355124

Shum, S. B., Liddo, A. D., Bachler, M., & Cornish, H. (2015, March 31). Debate Hub. Retrieved March 31, 2015, from http://debatehub.net/ui/pages/about.php

Siangliulue, P. (2015). Supporting Collaborative Innovation at Scale. In *Adjunct Proceedings of the 28th Annual ACM Symposium on User Interface Software & Technology* (pp. 9–12). New York, NY, USA: ACM. https://doi.org/10.1145/2815585.2815588

Siangliulue, P. (2017, April). *Supporting Collective Ideation at Scale*. Harvard University, Cambridge, MA. Retrieved from http://scs.cmu.edu/calendar/tue-2017-04-18-1330/crowdsourcing-lunch-seminar

Siangliulue, P., Arnold, K. C., Gajos, K. Z., & Dow, S. P. (2015). Toward Collaborative Ideation at Scale: Leveraging Ideas from Others to Generate More Creative and Diverse Ideas. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing* (pp. 937–945). New York, NY, USA: ACM. https://doi.org/10.1145/2675133.2675239

Simon, H. A. (1996). *The Sciences of the Artificial*. MIT Press.

Sinai, N., & Smith, G. (2013, December 6). The United States Releases its Second Open Government National Action Plan. Retrieved April 19, 2015, from http://www.whitehouse.gov/blog/2013/12/06/united-states-releases-its-second-open-government-national-action-plan

Singh, A., P, D., & Raghu, D. (2012). Retrieving Similar Discussion Forum Threads: A Structure Based Approach. In *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 135–144). New York, NY, USA: ACM. https://doi.org/10.1145/2348283.2348305

Sizov, S. (2012). Latent Geospatial Semantics of Social Media. *ACM Trans. Intell. Syst. Technol.*, *3*(4), 64:1–64:20. https://doi.org/10.1145/2337542.2337549

Sobel, D. (2005). *Longitude: The True Story of a Lone Genius Who Solved the Greatest Scientific Problem of His Time*. Macmillan. Retrieved from http://books.google.com/books?id=0v6DBAAAQBAJ

Spector, P. E. (1992). *Summated Rating Scale Construction: An Introduction*. SAGE. Retrieved from http://srmo.sagepub.com/view/summated-rating-scale-construction/n1.xml

Spertus, E., Sahami, M., & Buyukkokten, O. (2005). Evaluating Similarity Measures: A Large-Scale Study in the Orkut Social Network. In *Proceedings of the Eleventh ACM SIGKDD International Conference on Knowledge Discovery in Data Mining* (pp. 678–684). New York, NY, USA: ACM. https://doi.org/10.1145/1081870.1081956

Steenkamp, J.-B. E. M. (1990). Conceptual Model of the Quality Perception Process. *Journal of Business Research*, *21*(4), 309–333. https://doi.org/10.1016/0148-2963(90)90019-A

Stein, J. (2016, August 28). How Trolls Are Ruining the Internet. *Time*. Retrieved from http://time.com/4457110/internet-trolls/

Steinfeld, N., Samuel-Azran, T., & Lev-On, A. (2016). User Comments and Public Opinion: Findings from an Eye-Tracking Experiment. *Computers in Human Behavior*, *61*, 63–72. https://doi.org/10.1016/j.chb.2016.03.004

Steyvers, M., & Griffiths, T. (2007). Probabilistic Topic Models. In T. K. Landauer, D. S. McNamara, S. Dennis, & W. Kintsch (Eds.), *Handbook of Latent Semantic Analysis* (pp. 427–448). Mahwah, N.J: Lawrence Erlbaum Associates. Retrieved from http://www.taylorandfrancis.com/books/details/9781138004191/

Stuart, H. C., Dabbish, L., Kiesler, S., Kinnaird, P., & Kang, R. (2012). Social transparency in networked information exchange: a theoretical framework. In *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work* (pp. 451–460). New York, NY, USA: ACM. https://doi.org/10.1145/2145204.2145275

Stvilia, B., Twidale, M. B., Smith, L. C., & Gasser, L. (2008). Information Quality Work Organization in Wikipedia. *Journal of the American Society for Information Science and Technology*, *59*(6), 983–1001. https://doi.org/10.1002/asi.20813

Sullivan, D. (2014, December 27). Yahoo Directory Closes, Five Days Early. Retrieved January 1, 2015, from http://searchengineland.com/yahoo-directory-closes-211784

Sunstein, C. R. (2006). *Infotopia: How Many Minds Produce Knowledge*. New York; Oxford: Oxford University Press.

Surowiecki, J. (2005). *The Wisdom of Crowds*. Random House, Inc.

The Open Government Partnership. (2013). *Second Open Government National Action Plan for the United States of America*. Washington, DC, USA: The White House. Retrieved from http://www.whitehouse.gov/sites/default/files/docs/us_national_action_plan_6p.pdf

The White House. (2012, May). Youth Sustainability Challenge. Retrieved April 20, 2015, from https://www.whitehouse.gov/sustainability-challenge

Thomas, K. W. (1992). Conflict and conflict management: Reflections and update. *Journal of Organizational Behavior*, *13*(3), 265–274. https://doi.org/10.1002/job.4030130307

Towne, W. B. (2009). From "Localized" to "Networked:" A Transformation in Community Structures. In *ISTAS '09* (pp. 1–7). Tempe, AZ. https://doi.org/10.1109/ISTAS.2009.5155909

Towne, W. B. (2014, August 20). "Green Bond" Crowdfunding for Solar Projects. Retrieved November 29, 2016, from http://climatecolab.org/plans/-/plans/contests/2014/energy-supply/c/proposal/1309349

Towne, W. B., & Herbsleb, J. D. (2012). Design Considerations for Online Deliberation Systems. *Journal of Information Technology & Politics*, *9*(1), 97–115. https://doi.org/10.1080/19331681.2011.637711

Towne, W. B., Kittur, A., Kinnaird, P., & Herbsleb, J. (2013). Your Process Is Showing: Controversy Management and Perceived Quality in Wikipedia. In *Proceedings of the 2013 Conference on Computer Supported Cooperative Work* (pp. 1059–1068). San Antonio, TX, USA: ACM. https://doi.org/10.1145/2441776.2441896

Towne, W. B., Rosé, C. P., & Herbsleb, J. D. (2016a). Measuring Similarity Similarly: LDA and Human Perception. *ACM Transactions on Intelligent Systems and Technology*, *8*(1), 7:1–7:28. https://doi.org/10.1145/2890510

Towne, W. B., Rosé, C. P., & Herbsleb, J. D. (2016b). The Key Role of Navigation in Online Challenge Platforms. Presented at the Collective Intelligence, New York, NY. Retrieved from https://sites.google.com/a/stern.nyu.edu/collective-intelligence-conference/

Towne, W. B., Rosé, C. P., & Herbsleb, J. D. (2017). Conflict in Comments: Learning but Lowering Perceptions, With Limits. In *CHI '17*. Denver, CO: ACM.

Traina, A. J. M., & Traina, C. (2017, May). *New Visions on Similarity Queries in DBMS: Big Data and NoSQL*. Presented at the Database Seminar, Carnegie Mellon School of Computer Science. Retrieved from http://scs.cmu.edu/calendar/tue-2017-05-02-1200/database-seminar

Trufelman, A. (2016, November). *Reverb: The Evolution of Architectural Acoustics*. Presented at the 99% Invisible. Retrieved from http://99percentinvisible.org/episode/reverb-evolution-architectural-acoustics/

Tsay, J., Dabbish, L., & Herbsleb, J. (2014). Let's Talk About It: Evaluating Contributions Through Discussion in GitHub. In *Proceedings of the 22nd ACM SIGSOFT International Symposium on Foundations of Software Engineering* (pp. 144–154). New York, NY, USA: ACM. https://doi.org/10.1145/2635868.2635882

Underwood, J. (2015, December 30). Curiosity: A Contextual Wikipedia Reader. Retrieved March 11, 2016, from https://www.macstories.net/reviews/curiosity-a-contextual-wikipedia-reader/

U.S. Department of Health & Human Services. (2013, October 9). Card Sorting. Retrieved May 4, 2015, from http://www.usability.gov/how-to-and-tools/methods/card-sorting.html

Valdés-Pérez, R. E. (1999). Discovery Tools for Science Apps. *Commun. ACM*, *42*(11), 37–41. https://doi.org/10.1145/319382.319389

Viégas, F. B., Wattenberg, M., & Dave, K. (2004). Studying Cooperation and Conflict Between Authors with History Flow Visualizations. In *Proceedings of the SIGCHI conference on Human factors in computing systems* (pp. 575–582). New York, NY, USA: ACM. https://doi.org/10.1145/985692.985765

Viégas, F. B., Wattenberg, M., Kriss, J., & van Ham, F. (2007). Talk Before You Type: Coordination in Wikipedia. In *Proc. HICSS 2007* (p. 78). https://doi.org/10.1109/HICSS.2007.511

Voorhees, E. M. (2007). TREC: Continuing Information Retrieval's Tradition of Experimentation. *Commun. ACM*, *50*(11), 51–54. https://doi.org/10.1145/1297797.1297822

Vozick, T. (2006, December 13). Consensus on intro wording in is Archive- NEW EDITORS PLEASE READ. In *Wikipedia*. Retrieved from https://en.wikipedia.org/w/index.php?title=Talk:Barack_Obama/Archive_4&oldid=390259634

Walter, T. P., & Back, A. (2013). A Text Mining Approach to Evaluate Submissions to Crowdsourcing Contests. In *2013 46th Hawaii International Conference on System Sciences (HICSS)* (pp. 3109–3118). https://doi.org/10.1109/HICSS.2013.64

Wang, Y., Joshi, M., Cohen, W., & Rosé, C. (2008). Recovering Implicit Thread Structure in Newsgroup Style Conversations. In *Proceedings of the 2nd International Conference on Weblogs and Social Media* (pp. 152–160). Seattle, WA: AAAI. Retrieved from http://www.aaai.org/Papers/ICWSM/2008/ICWSM08-026.pdf

Wang, Y.-C., Joshi, M., & Rosé, C. P. (2008). Investigating the Effect of Discussion Forum Interface Affordances on Patterns of Conversational Interactions. In *Proceedings of the 2008 ACM Conference on Computer Supported Cooperative Work* (pp. 555–558). New York, NY, USA: ACM. https://doi.org/10.1145/1460563.1460650

Webcredible. (2015, May 4). Card Sorting. Retrieved May 4, 2015, from http://www.webcredible.com/services/card-sorting.shtml

Wikimedia Foundation. (n.d.). MediaWiki Extension:ArticleFeedback. Retrieved March 26, 2012, from https://www.mediawiki.org/wiki/Extension:ArticleFeedback

Wikipedia community consensus. (2012a, May 9). Wikipedia: Neutral Point Of View. In *Wikipedia, the free encyclopedia*. Wikimedia Foundation, Inc. Retrieved from https://en.wikipedia.org/w/index.php?title=Wikipedia:Neutral_point_of_view&oldid=491462094

Wikipedia community consensus. (2012b, May 26). Wikipedia: There Is No Deadline. In *Wikipedia, the free encyclopedia*. Wikimedia Foundation, Inc. Retrieved from https://en.wikipedia.org/w/index.php?title=Wikipedia:There_is_no_deadline&oldid=487857034

Wikipedia contributors. (2012, May 9). Help: Using Talk Pages. In *Wikipedia, the free encyclopedia*. Wikimedia Foundation, Inc. Retrieved from https://en.wikipedia.org/w/index.php?title=Help:Using_talk_pages&oldid=487941001

Wilson, T. D., Lisle, D. J., Kraft, D., & Wetzel, C. G. (1989). Preferences as Expectation-Driven Inferences: Effects of Affective Expectations on Affective Experience. *Journal of Personality and Social Psychology*, *56*(4), 519–530. https://doi.org/10.1037/0022-3514.56.4.519

Wolfers, J., & Zitzewitz, E. (2004). Prediction Markets. *Journal of Economic Perspectives*, *18*(2), 107–126.

Wolfers, J., & Zitzewitz, E. (2008, February 28). Prediction Markets. In S. N. Durlauf & L. E. Blume (Eds.), *The New Palgrave Dictionary of Economics* (2nd ed.). Basingstoke: Nature Publishing Group. Retrieved from http://www.dictionaryofeconomics.com/article?id=pde2008_P000340&goto=predictionmarkets&result_number=2841

Wolfram Alpha. (2016, April 27). Lee (given name). Retrieved April 27, 2016, from http://www.wolframalpha.com/input/?i=Lee+(given+name)

Wong, N. (2014, January). *Data Privacy Day Keynote*. Presented at the Data Privacy Day, Pittsburgh, PA. Retrieved from http://www.cs.cmu.edu/news/white-house-privacy-officer-keynote-speaker-carnegie-mellons-data-privacy-day-program-jan-29

Wong, Y. Y. (1992). Rough and Ready Prototypes: Lessons from Graphic Design. In *Posters and short talks of the 1992 SIGCHI conference on Human factors in computing systems* (pp. 83–84). New York, NY, USA: ACM. https://doi.org/10.1145/1125021.1125094

Woolley, A. W., Chabris, C. F., Pentland, A., Hashmi, N., & Malone, T. W. (2010). Evidence for a Collective Intelligence Factor in the Performance of Human Groups. *Science*, *330*(6004), 686–688. https://doi.org/10.1126/science.1193147

Word, D. L., Coleman, C. D., Nunziata, R., & Kominski, R. (2016). *Frequently Occurring Surnames Data*. US Census Bureau. Retrieved from http://www.census.gov/topics/population/genealogy/data.html

Xu, G., & Ma, W.-Y. (2006). Building Implicit Links from Content for Forum Search. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 300–307). New York, NY, USA: ACM. https://doi.org/10.1145/1148170.1148224

Yin, R. K. (2013). *Case Study Research: Design and Methods* (5th ed.). Thousand Oaks, CA: SAGE Publications. Retrieved from https://books.google.com/books?id=OgyqBAAAQBAJ

Yin, Z., Cao, L., Gu, Q., & Han, J. (2012). Latent Community Topic Analysis: Integration of Community Discovery with Topic Modeling. *ACM Trans. Intell. Syst. Technol.*, *3*(4), 63:1–63:21. https://doi.org/10.1145/2337542.2337548

Yu, B., Kaufmann, S., & Diermeier, D. (2008). Exploring the Characteristics of Opinion Expressions for Political Opinion Classification. In *Proceedings of the 2008 International Conference on Digital Government Research* (pp. 82–91). Montreal, Canada: Digital Government Society of North America. Retrieved from http://dl.acm.org/citation.cfm?id=1367832.1367848

Zane, J. P. (2012, November 8). MacArthur Fellows Imagine Prizes to Spur Problem-Solving. *The New York Times*. Retrieved from http://www.nytimes.com/2012/11/09/giving/macarthur-fellows-imagine-prizes-to-spur-problem-solving.html

Zhai, Z., Liu, B., Xu, H., & Jia, P. (2011). Constrained LDA for Grouping Product Features in Opinion Mining. In J. Z. Huang, L. Cao, & J. Srivastava (Eds.), *Advances in Knowledge Discovery and Data Mining* (pp. 448–459). Springer Berlin Heidelberg. Retrieved from http://link.springer.com/chapter/10.1007/978-3-642-20841-6_37

Zhang, H., Qiu, B., Giles, C. L., Foley, H. C., & Yen, J. (2007). An LDA-based Community Structure Discovery Approach for Large-Scale Social Networks. In *2007 IEEE Intelligence and Security Informatics* (pp. 200–207). https://doi.org/10.1109/ISI.2007.379553

Zhao, S., Zhou, M. X., Zhang, X., Yuan, Q., Zheng, W., & Fu, R. (2011). Who is Doing What and When: Social Map-Based Recommendation for Content-Centric Social Web Sites. *ACM Trans. Intell. Syst. Technol.*, *3*(1), 5:1–5:23. https://doi.org/10.1145/2036264.2036269

(1869, March 29). *Daily Cleveland Herald*.