Supporting Effective Collective Ideation at Scale

a dissertation presented

by

Kanya (Pao) Siangliulue

to

The School of Engineering and Applied Sciences

in partial fulfillment of the requirements

for the degree of

Doctor of Philosophy

in the subject of

Computer Science

Harvard University

Cambridge, Massachusetts

May, 2017

Dissertation advisor: Krzysztof Z. Gajos Kanya (Pao) Siangliulue

Supporting Effective Collective Ideation at Scale

Abstract

Online collective ideation platforms, such as OpenIDEO or Quirky, have demonstrated the potential of large-scale collective innovation in various domains. However, the users of these platforms face new challenges of leveraging collective contributions. The large number of collected ideas prevents users from making full use of these ideas. Finding inspirations from the ideas involves wading through a sea of possibly mundane and redundant ideas. Synthesizing a few solutions from these ideas takes a lot of time and effort. I argue that leaving users to explore ideas in a haphazard manner is ineffective and can decrease the quality of people's creative output. Prior work in cognitive science and creativity research has also suggested that deliberate exploration of the solution space can improve users' creative output and experience.

I introduce the concept of an *idea map*, a computational model of the emerging solution space that enables deliberate exploration interactions: 1) presenting a set of ideas with a controlled level of diversity appropriate to the stage of the creative process and 2) presenting a summary view of the solution space. I describe two scalable crowdsourced methods for generating this computational model. The first method computes the model from responses from small micro-task questions. The second method takes an "integrated crowdsourcing" approach that computes the model from users' natural activities during idea generation. The evaluation of the derived models show that the idea maps from both approaches agree with human judgments of similarities among ideas. I show the application of the idea map concept through experiments and a system called IdeaHound. IdeaHound derives an idea map using the integrated crowdsourcing approach and uses the derived model to guide

users' exploration of the solution space. The results of the experiments show that an idea map can inspire people to generate diverse ideas. The integrated activities that enable IdeaHound to collect similarity judgments do not deter users from generating ideas and provide enough information to generate a reliable idea map. I also present a study on the effects of different timings of delivering example ideas on an individual's idea generation. The results demonstrate that an intelligent system can provide inspiration at the right moment by using a computational model that is aware of semantic relationships between ideas. Finally, I demonstrate how to use an idea map to support sensemaking during the solution synthesis and present an empirical study of the effect of presenting a summary view of ideas on people's solution synthesis.

Contents

Citations to Previously Published Work

Significant portions of this dissertation work have appeared in published papers.

Chapter 3 has adapted, updated, and rewritten content from the following paper:

Siangliulue, Pao, Kenneth C. Arnold, Krzysztof Z. Gajos, and Steven P. Dow. "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. ACM, 2015.

Chapter 4 has adapted, updated and rewritten content from the following paper:

Siangliulue, Pao, Joel Chan, Bernd Huber, Steven P. Dow, and Krzysztof Z. Gajos. "IdeaHound: Self-sustainable Idea Generation in Creative Online Communities." In Proceedings of the 19th ACM Conference on Computer Supported Cooperative Work and Social Computing Companion, pp. 98-101. ACM, 2016.

Chapter 5 has adapted, updated, and rewritten content from the following paper:

Siangliulue, Pao, Joel Chan, Krzysztof Z. Gajos, and Steven P. Dow. "Providing timely examples improves the quantity and quality of generated ideas." In Proceedings of the 2015 ACM SIGCHI Conference on Creativity and Cognition, pp. 83-92. ACM, 2015.

Acknowledgments

I am grateful for the opportunity to deeply pursue my research interest as a PhD student. I owe this fulfilling graduate school career to many.

First, I want to thank my advisor Krzysztof Gajos for his insights, mentorship, and support. Krzysztof taught me to articulate my thoughts in a precise manner and pursue answers through rigorous methods. He encouraged me to pursue hard questions and offer advice when I was stuck. His patience and guidance gave me the necessary space to explore and grow as an independent researcher and a person. It is not an exaggeration to say that I would not have completed this dissertation without his tutelage.

Steven Dow has been a constant influence throughout my graduate school career and even before. His expertise and creativity in framing research questions and designing an experiment contributed tremendously to this dissertation. His feedback usually opened new directions or questions that I have not thought of before.

Barbara Grosz gave me great feedback while writing this dissertation. She provided me with prudent advice on framing my research and career during and beyond my graduate school. I am also grateful for her occasional chocolate donation to our office that kept me awake while working late at night.

I am fortunate to have a chance to work with incredible collaborators. I thank Joel Chan for his knowledge on creativity research, clear insights on crucial experimental design and positive energy he brought to the team. I thank Ken Arnold for his machine learning insights that formed the concept of an *idea map* and helpful feedback on following work. I thank Bernd Huber for his contribution in developing and demoing IdeaHound. I thank my mentors Holger Winnemoeller, Mira Dontcheva and Sheryl Ehrlich during my internship at Adobe for their guidance and an opportunity to expand my research interests and experiences.

The conversations with Michael Terry, Andrew Mao, Alice Gao, and Yuval Hart helped provide me with backgrounds and new thoughts about my research in creativity support tools. I also thank participants in my studies whose opinions and feedback improved my understanding of people.

A big thank you to the IIS group members, alumni, visitors and my office-mates for insightful feedbacks, advices, friendship, fun time and a welcoming space to grow: Ken Arnold, Bernd Huber, Ofra Amir, Sebastian Gehrmann, Steve Komarov, Anna Huang, Elena Agapie, Katharina Reinecke, Sunyoung Kim, Na Li, Joseph Williams, Juho Kim, Tuukka Ruotsalo, Nam Wook Kim, Dimitrios Antos, and Hsiao. Also special thanks to admins Shannon Cardillo, David Lopez, Jess Jackson for keeping things running smoothly.

Thanks to the Stanford HCI group during my undergraduate years, especially Scott Klemmer for igniting my interest in the HCI field and graciously providing me with an opportunity to try my hand at HCI research.

I would also like to thank my friends in both Cambridge and beyond for keeping my sanity. I thank N'May and Jao for fun conversations and karaoke nights. I am also grateful for the Harvard Aikikai community for providing me with a healthy way to decompress and enjoy good company.

Lastly, I want to thank my family: mom, dad and P'Booey. I thank my dad for fond memories and wise words that still reach me years after he passed. I thank my big sister P'Booey for her gentle care, support and for being the responsible one. I thank my mom for her love, encouragement, and cross-continent wake-up calls.

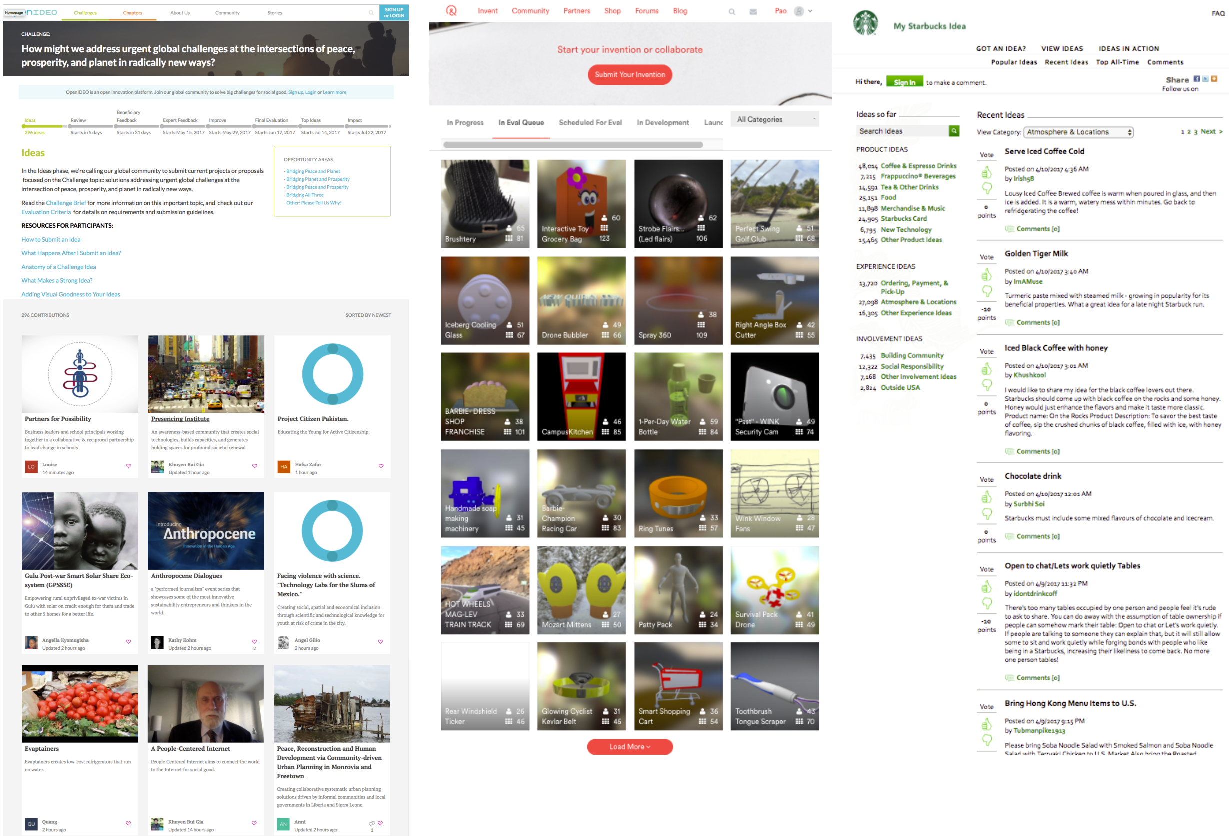
This dissertation work is supported in parts by gifts from Adobe and Google and the Siebel Scholars award.

Chapter 1

Introduction

Large collective ideation platforms have potential to transform the way our society innovates. Existing platforms have already attracted a large number of people to contribute ideas for problems in various domains: OpenIDEO has more than 100,000 innovators who have submitted thousands of ideas to solve social problems; Quirky, a collaborative invention platform, has built a community of more than 1,000 inventors who have proposed more than 10,000 product ideas; the City of Cambridge's Participatory Budgeting asked Cambridge residents to brainstorm ideas to improve the city resulting in 43 projects—on more than a million dollar budget—synthesized from 1,408 ideas. With these platforms and communities, it is possible for anyone to contribute ideas for problems they care about. We can now leverage this immense number and diversity of experiences and perspectives that lead to more diverse and creative solutions than ever before possible.

Creativity has different definitions from various traditions of research. In this dissertation, I am using a sociocultural definition which defines creativity as "*the generation of a product that is judged to be novel and also to be appropriate, useful, or valuable by a suitably knowledgeable social group*" [Sawyer, 2011]. Creative ideas are generally products of combinations between thoughts and concepts [Finke et al., 1992, Sawyer, 2011]. Large collective ideation platforms



(A) OpenIDEO (B) Quirky (C) My Starbucks Ideas

Figure 1.1: Many current platforms—for example, (A) Openideo.com, (B) quirky.com, and (C) mystarbucksidea.force.com—present ideas in a big list that can be sorted according to how popular and recent they are. Users have to browse through a lot ideas to find ones that inspire them.

provide more raw materials in the form of ideas of others that inspire an individual to generate unexpected ideas and can thus increase the chance of producing creative solutions overall.

However, current platforms and systems lack tools to help people—either contributors or organizers—make full use of this scale of contribution. Figure 1.1 shows how ideas are presented in existing platforms. To seek inspirations from ideas of others, users have to wade through a sea of ideas that can potentially harm their creative output [Jansson and Smith, 1991a, Kohn and Smith, 2011] because a lot of collected ideas are simple, mundane and repetitive [Bjelland and Wood, 2008]. Further, the large number of redundant and mundane ideas hinders solution synthesis after all ideas are collected. Currently, synthesizers (e.g., challenge organizers, decision makers, stakeholders, and communities representatives) have to read through all ideas and evaluate them one by one before synthesizing them into viable solutions. The synthesis process takes a lot of time and effort, which increases with the number of collected ideas. In some crowd innovation challenges, this process can take up to a few months [Klein and Garcia, 2015].

Like any other creative process, the creative process of large scale ideation can be divided into two alternating phases: the divergent phase and the convergent phases [Laseau, 2001, Buxton, 2007]. During the divergent phase, people aim to generate a lot of creative and diverse ideas. During the convergent phase, people synthesize the large collection of ideas into a small number of representative solutions.

Prior research in creative cognition and sensemaking indicates that people perform better when given guidance while looking for inspirations in the divergent phase and synthesizing solutions in the convergent phase, compared to when they were left to explore ideas in a haphazard manner. Specifically, previous work on creative cognition suggests that, during the divergent phase, people generate more creative and diverse ideas when they are exposed to creative and diverse examples [Nijstad et al., 2002, Nijstad and Stroebe, 2006, Siangliulue et al., 2015a]. These insights can inform the design of an intelligent system that reasons about when and what to present to the users to best inspire them. Also, previous work on sensemaking suggests that during the convergent phase, people can more quickly develop a deep understanding of a dataset if they are provided with an initial schema or summary of the data instead of having to spend a lot of time extracting schema from a lot of ideas [Fisher et al., 2012, Kittur et al., 2014]. It is thus my thesis that:

In large-scale collective ideation settings, intelligent systems that understand the emerging solution space of ideas can improve users' creative output by 1) recommending sets of inspiring ideas during the divergent phase and 2) providing

an interpretable summary of idea space during the convergent phase of ideation.

In this dissertation, I introduce the concept of an *idea map*, a computational model of an emerging solution space. An idea map approximates similarities among ideas which could help a system selects a set of diverse ideas. I describe two scalable approaches for computing an idea map using crowdsourcing and machine learning techniques. The first approach generates an idea map from human responses to micro-task questions about relationships among ideas. The second approach derives an idea map from users' natural activities during idea generation without needing additional human inputs. I will refer to the second approach as "integrated crowdsourcing". I present studies that evaluate the reliability of idea maps generated by these two approaches. The results show that idea maps generated by both approaches agree with human judgments of similarity between pairs of ideas; an idea map derived from users' natural activities agreed more with human judgments than an idea map generated from micro-task responses. My results also show that presenting people with a diverse set of ideas sampled from an idea map inspires diverse ideas. I demonstrate ways an idea map can enable deliberate exploration of a solution space with IdeaHound, a system that derives an idea map from the users' collective natural activities of organizing ideas on a virtual whiteboard during idea generation. I also present a study of the effects of timing of example delivery on people's creative output to determine when to provide people with ideas. Finally, I describe a study of the effect of seeing a summary view of ideas on people's solution synthesis. The results of this study demonstrates that a summary view that groups similar ideas helps people spot rare ideas but can fixate people on the single point of view the summary suggests.

The following sections summarize the contributions of this dissertation.

1.1 Idea Map: Scalable Dynamic Model of Idea Space

Ideas submitted by other people can serve as an important resource for creative production but they have to be used with care. Seeing inspiring ideas can help improve creative output [Herring et al., 2009, Marsh et al., 1996], while seeing uninspiring ideas can cause fixation on those ideas [Chrysikou and Weisberg, 2005, Jansson and Smith, 1991b]. Many factors determine whether an idea is inspiring for an individual. Research has considered such factors as semantic relevance [Chan et al., 2014, 2011, Dahl and Moreau, 2002], novelty [Chan et al., 2011, Agogué et al., 2013], and diversity [Doboli et al., 2014, Zeng et al., 2011, Baruah and Paulus, 2011, Siangliulue et al., 2015a] of example ideas. Prior research suggests that presenting users with inspirational examples that are both creative and diverse increases the creativity and diversity of generated ideas [Marsh et al., 1996, Paulus and Dzindolet, 1993, Nijstad et al., 2002, Leggett Dugosh and Paulus, 2005]. While a few scalable approaches to selecting creative ideas exist [Xu and Bailey, 2012, Klein and Garcia, 2015], methods for selecting diverse sets of ideas are less well-developed.

To select inspiring ideas from a large collection in real-time as new ideas come in, the systems need an approach that is both scalable (to handle a large number of ideas) and dynamic (to integrate new ideas as examples almost instantly instead of waiting for batch processing). Fully automated methods that satisfy both requirements currently fall short when processing ideas that are less structured and expressed in the form of short text snippets or sketches [Chang et al., 2009, Talton et al., 2009, Lee et al., 2010, Gerber et al., 2012]. Popular automated text processing approaches such as Latent Dirichlet allocation (LDA) do not always infer semantically meaningful topics [Chang et al., 2009, Blei, 2012]. Approaches that incorporate human inputs are more flexible and capture non-superficial similarity [Chang et al., 2009, André et al., 2014]. However, prior human-computation approaches either require a lot of human input to account for similarities between all pairs of ideas or process ideas in batches rather than incrementally.

In Chapter 3, I present a scalable crowd-powered approach for selecting a set of diverse ideas. This approach, adapted from an existing machine-learning method [Tamuz et al., 2011], uses similarity comparisons. It asks non-experts "is idea A more similar to B or C?" and uses their responses to generate a computational model called an *idea map* (Figure 3.1). An idea map encodes similarities between ideas. It is an approximation of a full pairwise distance matrix. An idea map requires as few as $O(n)$ distance comparisons as input (compared to $O(n^2)$ needed to compute the full distance matrix) and infers the rest by embedding ideas in a two-dimensional space that is most consistent with the similarity comparisons people made. On the idea map, similar ideas are placed close to each other and dissimilar ideas are placed far from each other. An ideation system can algorithmically select a set of diverse ideas by sampling ideas from different parts of the idea map.

The evaluation of a generated idea map reveals that the idea map reflects people's judgments of conceptual relationships among ideas. Human raters agree with the estimates of similarity of derived from the idea map as much or more than they agree with each other. Further, presenting a diverse set of ideas sampled from an idea map prompts people to generate more diverse ideas than when they see a randomly sample set of ideas.

1.2 IdeaHound: Integrated Crowdsourcing Approach for Creativity Enhancing Interventions

An idea map can enable creative interventions that boost creativity, such as providing diverse inspirational examples. However, generating an idea map from micro-task responses requires extra human effort. Even people who might be intrinsically motivated to generate ideas are unlikely to put effort toward completing the micro-tasks. These micro-taskes are usually outsourced to external workers who complete a large number of them. However, outsourcing such tedious tasks is not always a viable solution. Furthermore, some types of ideas, such as scientific work, require expertise to process and evaluate. Untrained crowd workers are thus not qualified to evaluate the ideas and outsourcing such micro-tasks to skilled workers is expensive or impossible.

Chapter 4 introduces a general approach called "integrated crowdsourcing" that seamlessly integrates a potentially tedious secondary task with a more intrinsically-motivated primary task. I demonstrate the application of this approach with IdeaHound, an ideation system that integrates the task of extracting similarities among ideas into the primary task of idea generation. The system uses the generated idea map to support three creative interventions: providing diverse inspirational examples, providing ideas similar to a given idea and providing a summary of the solution space.

IdeaHound provides users with a virtual whiteboard that they can use to organize ideas into groups, an activity that users naturally do when they are making sense of the solution space while generating ideas. The system infers similarity between ideas based on how users group them. IdeaHound combines the results of these implicit human actions with machine learning techniques to create an idea map. The integrated nature of the similarity extraction task allows IdeaHound to leverage the expertise and efforts of users who are already motivated to contribute to idea generation, overcoming the scalability limitations of existing approaches.

The evaluation of IdeaHound shows that the integrated task does not detract users from the main task of generating ideas and that the derived computational model is more accurate than a comparable model generated by an outsourced micro-task approach. Specifically, participants were equally willing to use IdeaHound compared to a conventional platform that did not require organizing ideas. These results show a successful application of integrated crowdsourcing on ideation platforms and promise further application of integrated crowdsourcing in other domains.

1.3 Timing of Example Delivery

Prior work has suggested the timing of delivery of examples can impact creative outcome [Kulkarni et al., 2014] but much less is known about *when* to present examples to people. Existing cognitive theories of creative insights suggest that people are likely to benefit most from examples when they run out of ideas [Seifert et al., 1995, Patalano and Seifert, 1994, Moss et al., 2007]. In Chapter 5, I present a study that explored two mechanisms that deliver examples when users are likely to be stuck: 1) a system that proactively provides examples when a user appears to have run out of ideas, and 2) a system that provides examples when a user explicitly requests them. The study compared these two mechanisms against two baselines: providing no examples and automatically showing examples at a regular interval.

The results show that participants who requested the examples themselves generated ideas that were rated the most novel by external evaluators; participants who received examples automatically when they appeared to run out of ideas produced the most ideas. Importantly, participants who received examples at a regular interval generated fewer ideas than participants who received no examples, suggesting that mere access to examples is not sufficient for creative inspiration.

These results help inform when to present inspiring ideas to users during ideation. The system should deliver examples when people are ready to make use of them, such as when they run out of ideas and are looking for a new direction to pursue. These results were further corroborated and extended by a recent study to which I contributed. That study demonstrated that receiving ideas that are different from people's recent ideas during productive ideation slows their ideation, reduce deep exploration of a topic and increase the chance of hitting an impasse [Chan et al., 2017].

1.4 Interpretable Summary Visualization for Solution Synthesis

The collective ideation process does not end at idea collection. Once the ideas are collected, synthesizers—usually the people who organize an ideation challenge or representatives of a community—have to synthesize them into a few different novel and practical solutions. This process involves manually looking through all ideas, comparing ideas against each other, evaluating ideas, and synthesizing solutions from multiple ideas. The process is laborious and time-consuming and may take months to complete [Klein and Garcia, 2015]. Furthermore, prior work in cognitive science [Nijstad and Stroebe, 2006, Finke et al., 1992] indicates that repeated exposure to the same ideas can fixate people to those particular ideas. The fixation effect could bias the synthesizers towards common ideas that cover the majority of the solution space.

Prior sensemaking research has shown that people can more quickly reach a deep understanding of a dataset if they are given an initial schema or "knowledge map" of the data [Fisher et al., 2012, Kittur et al., 2014]. An idea map already has information about conceptual similarities between ideas and thus can be used to generate a summary visualization of the emerging knowledge map of ideas since similar ideas are grouped together. This visualization shows each group of ideas in its own distinct cluster, regardless of the number of ideas in the group. A group with fewer ideas is thus as salient as a group with more ideas. I thus hypothesize that, with this summary view, users can also easily find rare ideas by looking at ideas that stand on their own or are in a small group.

However, reasoning over summaries of the solution space presents a potential trade-off between efficiency and effectiveness. For example, providing people with a summary visualization might speed up their solution synthesis process at the cost of getting fixated on the suggested schema. Ideas are multifaceted and a solution space can have multiple schemas.

Presenting just one schema to the users might prevent them from taking other schemas into account when they synthesize solutions. For instance, a synthesizer for a logo design might overlook schemas such as fonts and shapes of a logo when they see a summary view that groups logos by color scheme. Chapter 6 presents a study of this trade-off. The results show that the users with a summary view process more rare ideas and integrate more rare ideas into solutions than those without a summary view. However, users with a summary view also get more fixated on the schema suggested by the summary view. I discuss some approaches to mitigating the fixation on a particular schema while retaining the benefits of the summary view.

1.5 Overview

Chapter 2 describes major research challenges and related work.

The main part of the dissertation presents a study that informs effective example delivery mechanisms, a computational model that maps the solution space of a large collection of ideas, the evaluation of the model and two approaches to extract them, and two systems that apply the model to real creative task in collective ideation settings.

- Chapter 3 introduces a scalable crowdsourced approach that uses machine learning to generate a computational model, an *idea map*, of emerging ideas from micro-task inputs.
- Chapter 4 presents IdeaHound a system that derives an idea map from users' activities during idea generation. The system uses the generated idea map to support various creative interventions during ideation.
- Chapter 5 presents a study on the effects of timing of example delivery mechanisms on an individual's idea generation. The results inform design decisions on timing of example delivery for future systems.

- Chapter 6 describes a study on the effects of providing people with an interpretable summary view that shows ideas semantically grouped together on their solution synthesis.

Chapter 2

Related Work

This dissertation work is built upon work from various research areas: **creativity research** to provide insights on creativity enhancing interventions, **automated data-driven approaches** to inform the state-of-art of automatically uncovering a solution space of a large collection of ideas, **crowdsourcing and human computation** to efficiently harvest human perception of ideas where automated methods come short, and **sensemaking** to support the users' activities during the synthesis phase. In this chapter, I review related work in each area and situate it in the context of my research on collective ideation.

2.1 Creativity research

As noted in previous chapter, large-scale ideation systems typically do not live up to their promise in practice: they tend to collect large numbers of redundant and shallow ideas of variable quality [Bjelland and Wood, 2008, Klein and Garcia, 2015, Riedl et al., 2010]. The emerging literature on creative cognition and creativity support tools has identified a number of creativity-enhancing interventions that can significantly improve the performance of large-scale ideation systems by improving individual creativity and/or enhancing collaboration capabilities.

2.1.1 Cognitive Models of Creativity

Cognitive models of creativity have suggested that example ideas can have both positive effects and negative effects. For example, the model known as *Search for Ideas in Associative Memory* (SIAM) describes idea generation as a two-stage process: knowledge activation and idea production [Nijstad et al., 2002, Nijstad and Stroebe, 2006]. SIAM assumes two memory systems: long-term memory (permanent with unlimited capacity) and working memory (transient with limited capacity). Long term memory is partitioned into images, which are knowledge structures composed of a core concept and its features. For example, an image can have a core concept "hotel" with features like "has rooms", "has a swimming pool", and "is cheap". When generating ideas, people run a repeated two-stage search process. First, images from long term memory are retrieved and temporarily stored in working memory (knowledge activation). Then, in the second stage, the features of the image are used to generate ideas by combining knowledge, forming new associations, or applying them to a new domain (idea production). Retrieval of images probabilistically depends on search cues (e.g., features that are active in working memory, previously generated ideas, one's understanding of the problem). An image that is already in working memory is likely to be sampled again. SIAM, therefore, implies that seeing example ideas generally helps activate new images that would not have been accessible otherwise and thus leads to production of novel ideas. On the flip side, if the stimulus examples are homogenous, the generated ideas are likely to be homogenous, an exploration of semantically similar ideas in depth.

Similar to SIAM, the Geneplore model [Finke et al., 1992] also views examples as an activator of preinventive forms, a raw material for ideas in the exploration phase. If the set of stimulus examples is diverse, the generated ideas are likely to be diverse. This prediction from both models is supported by empirical evidence: people generated more diverse ideas when exposed to ideas from a wide range of semantic categories [Nijstad et al., 2002].

2.1.2 Timing of Inspiration Delivery

SIAM and the subsequent empirical results also indicate that example ideas can have both positive effects (cognitive stimulation) and negative effects (cognitive interference) based on **when** an example is shown [Nijstad et al., 2002, Nijstad and Stroebe, 2006]. On the one hand, seeing example ideas generally leads to production of novel ideas. On the other hand, ill-timed examples can prematurely terminate a person's train of thought, interrupt their thinking, and cause a loss of potentially creative ideas that usually come later in the session [Nijstad et al., 2002, Parnes, 1961, Bailey et al., 2000, Bailey and Iqbal, 2008].

Kulkarni et al. [Kulkarni et al., 2014] examined how the timing of examples affect creative output and concluded that early or repeated—rather than late—exposure to examples improves the creativity of generated ideas. However, Kulkarni et al. delivered examples at fixed regular intervals. This may not be optimal: intuitively, one might expect that people can be more or less "prepared" to benefit from examples at different points during the ideation process.

Beyond SIAM, several other theories of example use in problem solving and creative idea generation ground the intuition that people benefit more from examples when they are primed and ready. In education, the Preparation for Future Learning perspective [Schwartz and Martin, 2004, Schwartz et al., 2011] posits that learners get more out of learning resources (e.g., worked examples, lectures) if they first struggle with the concepts before being exposed to those resources. Relatedly, Kapur and colleagues have shown the value of "productive failure," a two-phase instructional strategy where students first engage in generation activities (e.g., attempting to solve problems that require knowledge of the target concepts) and then engage in consolidation/instruction, where they are exposed to the target concepts in various ways [Kapur, 2008]. These theories of learning argue that prior problem solving can prepare learners to let go of old knowledge, and prime them to notice important features of target concepts (e.g., what problem they are trying to solve).

The Prepared Mind theory of insight offers additional insights into the optimal timing of example idea presentation. It posits that people can be more or less "prepared" to assimilate problem-relevant stimuli from the environment depending on their cognitive state [Seifert et al., 1995, Patalano and Seifert, 1994]. The theory predicts specifically that, when problem solving reaches an impasse, people maintain an open goal in memory to solve the problem, and are more motivated and better able to map problem-relevant stimuli that might have been previously ignored (e.g., because it was too semantically distant or difficult to understand/transfer). Indeed, Tseng, et al. [Tseng et al., 2008] showed that people benefit more from analogically distant examples (a type of example hypothesized to be beneficial for creative inspiration [Dahl and Moreau, 2002]) during a break from problem solving after working on the problem for a significant amount of time compared to seeing the examples before working on the problem. Similarly, Moss, et al. [Moss et al., 2007] showed that people benefited more from hints after leaving a problem in an unsolved state compared to seeing the hints before working on the problem.

The shared intuition behind all of these theories is that optimal timing of example use for creative inspiration should strike a balance between allowing the ideator to queue up their own knowledge and constraints and avoiding cognitive fixation on a certain part of solution space. This intuition signals that delivering examples *when* people run out of ideas could maximize the inspirational benefit of examples. At that point, the examples can act as external stimuli to activate new knowledge in memory to combine into new ideas.

2.1.3 Creativity Enhancing Interventions

Seeing other people's ideas may have unintended side effects. Specifically, people tend to generate ideas that borrow concepts from presented examples [Jansson and Smith, 1991a, Smith et al., 1993, Marsh et al., 1996, Kohn and Smith, 2011]. If the examples were mundane or represented only a narrow slice of the solution space, seeing them may actually constrain rather than stimulate idea generation. This phenomenon has been referred to as *design fixation.*

While exposure to mundane examples may hinder creativity [Kohn and Smith, 2011, Jansson and Smith, 1991a], individuals can come up with more diverse and/or creative ideas if they have access to diverse and high quality inspirational examples [Chan et al., 2011, Nijstad et al., 2002, Marsh et al., 1996, Sio et al., 2015]. Empirical work suggests that exposing people to novel ideas, as opposed to common ones, can result in more novel ideas [Marsh et al., 1996]. Teams where members can see ideas of others generate more ideas (and sometimes of higher quality) compared to teams where each member generates ideas alone without seeing ideas of others [Gallupe et al., 1991, 1992, Dennis and Valacich, 1993]. Exposure to others' ideas also accelerates the generation of ideas across different semantic categories, increasing productivity overall [Nijstad et al., 2002]. This result might be attributable to the conforming effect: influenced by the novel examples, people incorporate the novel elements into their own ideas. An example with unfamiliar semantic properties prompt people to investigate ideas with those properties. Meanwhile, they might incorporate the ideas of their own with the examples, producing ideas in a new category that has not been explored by prior contributors. SIAM and the Geneplore model also suggests that people benefit from exposure to diverse sets of examples [Finke et al., 1992, Nijstad et al., 2002, Nijstad and Stroebe, 2006].

Another mechanism at play may be social influence. Results from a study of social influence processes in group brainstorming suggested that people are affected by information about the performance of others [Paulus and Dzindolet, 1993, Leggett Dugosh and Paulus, 2005]. One can infer the overall performance of others from ideas that one sees and try to match with ideas of the same caliber. Thus, exposing people to high quality, creative examples generated by peers can raise their aspirations while showing them less creative examples would likely lower the quality of subsequent ideas.

The net effect of increasing individual creativity is that the group can converge on novel, high quality solutions more quickly than if all participants simply saw their own ideas [Boudreau and Lakhani, 2015]. Inspirational examples can be drawn from peers' ideas for the same problem [Nijstad et al., 2002], or from external sources [Chan et al., 2011, Huang et al., 2016, Lee et al., 2010]. It is important that example sets be relatively small, because participants have limited time and cognitive resources [Javadi, 2012, Majchrzak and Malhotra, 2013]. If people have to process large numbers of examples, they can resort to effort-saving but suboptimal strategies, such as merely referring to (instead of deeply building on) other ideas [Javadi, 2012]. Further, the content of the ideas people see also matters. Prior research has found that examples are most inspirational if they are diverse [Huang et al., 2016] and/or appropriate to their current context [Huang et al., 2016]. Examples can also increase creativity by supporting exploration of iterations and variations on a solution approach [Sio et al., 2015, Chan and Schunn, 2015, Lee et al., 2010], which can lead to not just higher quality [Dow et al., 2010], but also more novel ideas [Nijstad et al., 2010a, Rietzschel et al., 2007b]. In contrast, poorly chosen examples can even harm ideation, by inducing distraction [Nijstad et al., 2002] or fixation [Kohn and Smith, 2011, Jansson and Smith, 1991a].

These insights point to two interventions that a collective ideation system might employ: show examples of particularly good ideas generated by others, and show a diverse sets of examples. In line with prior work, I hypothesize that both of these interventions will increase both the *creativity* and the *diversity* of ideas generated.

Individuals and groups can also achieve better creative outcomes if they have access to a "map" of the solution space that shows the kinds of solutions that have been explored by the group so far and/or in prior/external efforts to solve the problem, and how they relate to each other semantically. The map's higher-level view of the solution space can enable deeper insights into the solution space [Marks et al., 1997, Talton et al., 2009, Gerber et al., 2012], and the abstract solution "schemas" that might describe clusters of related ideas [Yu et al., 2014b]). These deeper insights have been shown to facilitate more effective recombination of ideas than with raw example ideas [Luo and Toubia, 2015, Yu et al., 2014b]. These maps can also improve iteration on ideas by enabling people to discover and explore many closely related solution alternatives [Lee et al., 2010, Chaudhuri et al., 2013, O'Donovan et al., 2014, Huang et al., 2014].

At the group level, maps have also been shown to help the group keep track of their exploration of the solution space [Nickerson et al., 2008, Nickerson and Yu, 2012]. The map can give participants an overall sense of what ideas have already been conceived, what "gaps" might exist, and where to focus their efforts. Participants can then make the best use of their limited time to make contributions that are most valuable to the group, avoiding redundant effort. This coordination benefit is supported by simulation studies [Wisdom and Goldstone, 2011, Vuculescu and Bergenholtz, 2014], as well as an empirical study of collaborative ideation on programming problems [Boudreau and Lakhani, 2015]. These maps greatly reduce the costs of manual coordination across collaborators, which can be extremely high in large-scale collaboration systems [Kittur et al., 2009]. However, these benefits have heretofore largely been realized in systems that engaged a small number of dedicated leaders to manually construct such maps.

The common thread behind these interventions is that they depend on having access to both a large corpus of solutions (whether generated externally or by peers in the same group) *and* a semantic model that specifies the structure of the solution space (e.g., how solutions relate to each other).

2.2 Automated Data-driven Approaches

Some automated mechanisms that extract information about the emerging solution space from a collection of ideas exist for domains where ideation artifacts are created with well-defined structures, such as webpages or geometric shapes in 3D modeling [Lee et al., 2010, Gerber et al., 2012, Marks et al., 1997, Talton et al., 2009]. The mechanisms rely on the structured nature of the artifacts and cannot generalize to nascent ideas most commonly found in the form of unstructured short text snippets or sketches. Fully automated topic analysis approaches exist for analyzing large collections of unstructured text [Blei, 2012]. However, these approaches often miss key nuances in the data [Chang et al., 2009, Chuang et al., 2012], and struggle with short text snippets. They also cannot handle unstructured sketches without some initial segmentation of sketches into reasonable "units" (analogous to words in topic modeling of texts).

2.3 Crowdsourcing and Human Computation

Information about ideas collected from people suits the nature of unstructured ideas better than information derived from automated methods. However, acquiring such information is expensive compared to automated methods. The challenge is therefore to collect human inputs accurately and efficiently.

Prior work has already produced a number of scalable mechanisms for evaluating the quality of individual ideas. Some of them are already used in existing online idea generation platforms. For example, Quirky.com and OpenIDEO.com have used simple binary voting mechanisms to identify promising ideas. AllOurIdeas.org finds top ideas by deriving ranks of ideas from users' ranking of pairs of ideas [Salganik and Levy, 2012]. In other works where more refined measures are needed, users rate creativity of ideas on different Likert scales [Yu and Nickerson, 2011, Tanaka et al., 2011]. Xu and Bailey demonstrated a mechanism that helps ensure that voting results from non-experts match those of experts by aggregating non-experts' ratings of subset of ideas [Xu and Bailey, 2012]. Scalably assessing semantic relationships among ideas, however, has not been as well studied.

Two approaches to quantifying semantic relationship are common in prior work: labeling items with semantic categories, or evaluating subjective similarity between items independent of semantic categories.

Semantic Categories Label

Manually created semantic categories have been used in prior research in creativity, either as ways to select a diverse set of ideas [Nijstad et al., 2002], or as a way to evaluate the diversity of creative artifacts [Goldenberg et al., 2013, Jansson and Smith, 1991a].

Efficient crowd-based mechanisms exist to label large collections of items with semantic categories or tags. Some take the approach of generating labels or tags for each individual item [Law and Von Ahn, 2009], while others produce hierarchical taxonomies capturing the semantic structure of the concepts represented in the item set [Chilton et al., 2013, 2014, André et al., 2014]. Differences in contributors' mental models have been a persistent difficulty in semantic categorization even for experts [Chilton et al., 2013]. A complete system for organizing ideas should include elements of both discrete semantic categories and continuous quantitative similarity.

Idea Similarity

An alternative approach to quantifying diversity is quantifying how items are *related*, such as evaluating the diversity of creative artifacts by collecting similarity judgements on a numerical scale between pairs of ideas [Dow et al., 2010]. However this approach requires on the order of the square of the number of ideas, making it less feasible for large idea collections. Moreover, accurate measures require that ratings be calibrated. Alternative approaches consider pairwise rankings, which ask evaluators to choose one pair of items over the other and thus do not require calibration.

Machine learning techniques can help scale human judgments by inferring a latent structure for the items, such as clusters [Heikinheimo and Ukkonen, 2013, Gomes et al., 2011] or a Euclidean space [Tamuz et al., 2011, van der Maaten and Weinberger, 2012]. Like the semantic categorization approaches, these approaches seek a compact representation of items rather than explicitly encoding all relationships, but the latent structure has no intrinsic semantics.

A promising alternative leverages human computation, whether exclusively or in a hybrid system with machine intelligence [André et al., 2014, Chilton et al., 2014]. Human computation approaches have been successfully applied to organize artifacts in various domains such as 3D modeling [Talton et al., 2009, Chaudhuri et al., 2013], graphic designs [O'Donovan et al., 2015, 2014] and music composition [Huang et al., 2014]. These approaches all require considerable number of inputs from humans to discover the design space of ideas; some of these inputs are extracted from users' interactions with the system [Talton et al., 2009, Huang et al., 2014], but most of these inputs are from small, explicit human computation tasks, such as clustering subsets of items [André et al., 2014, Chilton et al., 2014], completing similarity comparisons between items [Tamuz et al., 2011], or identifying attributes of items [Chaudhuri et al., 2013, O'Donovan et al., 2015, 2014].

One key disadvantage of human computation micro tasks is that they tend to be uninteresting and repetitive. The specific activities (tagging, judgements of relative similarity) take time, do not directly contribute to the ideation process, and are often perceived as tedious. Contributors to the online platforms generally avoid doing tedious maintenance tasks (in this case providing information about ideas) to do more interesting tasks (generating ideas) [Kraut et al., 2012]. One could argue that these activities could be performed by external crowds hired specifically for the purpose. However, even if cost was not an issue, this approach can be challenging for those creative tasks where specialized domain knowledge is required.

Further, outsourcing is typically done in batches, but some of the key creativity-enhancing interventions of collaborative ideation systems (e.g., coordination via a solution space overview) require (near) real-time continuous updates to the model. My research seeks to increase the feasibility of human computation approaches by exploring ways to integrate the semantic organization tasks into participants' primary activities.

2.4 Sensemaking

Currently, synthesizing solutions from a collection of ideas involves processing idea one by one manually to understand the emerging solution space [Schulze et al., 2012]. With a lot of mundane and repetitive ideas, synthesizing solutions becomes tiresome [Klein and Garcia, 2015, Bjelland and Wood, 2008]. Existing methods propose scalable ways to filter ideas based on quality but did not provide a way to help people make senses of the solution space.

Sensemaking is an iterative process of searching for a representation and encoding information to achieve a certain goal [Russell et al., 1993]. Prior work has explored different approach in supporting making sense of a large amount of information. Sensemaker provides users with interfaces that help them explore data from heterogeneous sources and develop mental model based on the information context [Baldonado and Winograd, 1997]. Grokker2 helps users make sense of large document collections by presenting document in groups based an automated clustering algorithms [Russell et al., 2006]. Terveen et al. [1999] presents users with a summary visualization that groups similar websites together. Relying on similar network data sructure, Apolo provides users an interactive summary view that shows similar items grouped together based on the users' evolving mental model [Chau et al., 2011]. Alternatively, some prior work explores a summary view based on attributes of the information [Kittur et al., 2014, Russell et al., 1993].

Prior work on idea generation also explores different summary views to make sense of a collection of ideas. Idea Spotter shows a summary of individual ideas marked by other users[Convertino et al., 2013]. IdeaGens shows a word cloud visualization of submitted ideas to summarize evolving solution space to support facilitating synchronize ideation [Chan et al., 2016]. Both systems are limited to ideas in the form of text and did not support sensemaking by grouping similar ideas together. Prior work also explores different ways to synthesize insights for a design team and proposes showing summary views with similar ideas placed together [Gumienny et al., 2014].

Chapter 3

Idea Map: Semantic Model of Idea Space

This chapter has adapted, updated, and rewritten content from a paper at CSCW 2015 [Siangliulue et al., 2015b]. All uses of "we", "our", and "us" in this chapter refer to coauthors of the aforementioned paper.

A growing number of large collaborative idea generation platforms promise that by generating ideas together, people can create better ideas than any would have alone. But how might these platforms best leverage the number and diversity of contributors to help each contributor generate even better ideas?

Prior research suggests that seeing particularly creative or diverse ideas from others can inspire you, but few scalable mechanisms exist to assess diversity. This chapter introduces a new scalable crowd-powered method for evaluating the diversity of sets of ideas. The method relies on similarity comparisons (is idea A more similar to B or C?) generated by non-experts to create an abstract spatial *idea map*. Our validation study reveals that human raters agree with the estimates of dissimilarity derived from our idea map as much or more than they agree with each other. People seeing the diverse sets of examples from our idea map generate more diverse ideas than those seeing randomly selected examples. Our results also corroborate findings from prior research showing that people presented with creative examples generated more creative ideas than those who saw a set of random examples. We see this work as a step toward building more effective online systems for supporting large scale collective ideation.

3.1 Motivation and Contributions

The "lone inventor" is a myth: even geniuses benefit from exposure to ideas of others [Singh and Fleming, 2010]. Seeing ideas different from their own broadens people's perspectives, sheds light on obscure connections, and inspires people to come up with ideas they might not have thought of alone [Herring et al., 2009, Lee et al., 2010, Dow et al., 2011]. By generating ideas together, people can produce more diverse ideas than if each person generate ideas alone, and this diversity can lead to more creative overall solutions [Nagasundaram and Dennis, 1993, Paulus et al., 2011].

Various online platforms have emerged as spaces where people can share their ideas and get inspired by other people's ideas. For example, AllOurIdeas.org hosts more than 200,000 ideas addressing 4,500 problems, Quirky.com receives hundreds of new product ideas every day from its 500,000 inventors, and OpenIDEO.com hosts an archive of more than 1,000 ideas to solve 24 pertinent societal problems. Contributors to these platforms can browse other people's ideas in search of inspiration. The mix of perspectives and expertise among the participants allows creative solutions to emerge in a way unimaginable in the lone-innovator or small-group settings.

But the large-scale idea generation paradigm also introduces a new challenge: how to find the most inspiring ideas in a sea of hundreds? Existing approaches are to either help people parametrically browse and search for examples [Lee et al., 2010, Kumar et al., 2013] or extract schemas from examples and search for the schema that allows analogical transfer for a new idea [Yu et al., 2014a,b]. Even with such strategies, the users still have to wade through many examples to either find an inspiring idea or to find the right set of ideas to allow schema induction. They may get overwhelmed by a large number of mundane or redundant ideas before they encounter ideas that genuinely inspire them.

Alternatively, a system can select appropriate sets of inspiring examples for its users. The challenges of algorithmically identifying inspiring ideas from a large pool of ideas are twofold.

Firstly, picking out a set of inspiring ideas calls for finesse. People are easily influenced by ideas they encounter [Jansson and Smith, 1991a, Smith et al., 1993, Marsh et al., 1996, Kohn and Smith, 2011]. A set of uninspiring examples may fixate contributors on ordinary or a relatively narrow set of ideas. In contrast, a set of unique examples might prompt people to explore semantically different paths from their original ones, potentially yielding ideas from unexplored parts of the solution space. Our literature review reveals two criteria for an inspiring set of example ideas: creativity of individual examples and diversity of the *set* of examples. Compared to seeing a random selection of examples (as one might see when simply browsing ideas), seeing particularly creative (i.e., novel and valuable) ideation examples has been shown to improve both the creativity and diversity of ideas one generates [Marsh et al., 1996, Paulus and Dzindolet, 1993, Leggett Dugosh and Paulus, 2005]. Similarly, if the set of examples is diverse (i.e., if the ideas within the set were substantially different from each other), the diversity of generated ideas should also increase [Nijstad et al., 2002]. An effective ideation system should be able to assess the creativity of ideas and diversity of sets of ideas to be able to present inspiring ideas to promote creativity.

Secondly, there is a question of scalability. Even a human expert might struggle to find a set of creative and diverse ideas from a large idea archive in a reasonable amount of time. Our approach needs to effectively identify a promising set of inspiration from a pool of thousands of ideas.

Scalable crowd-powered mechanisms for assessing creativity of individual ideas have already been developed [Salganik and Levy, 2012, Tanaka et al., 2011, Xu and Bailey, 2012, Yu and Nickerson, 2011, Yu et al., 2014b,a]. However, automated or crowd-powered methods for assessing semantic diversity of sets of ideas are less well developed. To enable selection of diverse sets, we built on prior work on multidimensional scaling and active similarity learning techniques [Tamuz et al., 2011, van der Maaten and Weinberger, 2012] to develop a technique that "embeds" all ideas in a two-dimensional space, creating an abstract spatial map from as few human queries as possible. As input, our technique takes triplet comparisons ("Is idea A more similar to B or to C?"), which non-experts can provide easily and reliably. The distance between a pair of ideas on the generated map reflects the collective perception of the semantic difference between these ideas. This map allows us to estimate the relative diversity of subsets of ideas: sets where all ideas are close to each other on the map will be perceived as less diverse than sets where ideas are far apart.

We conducted a study to test whether the effects of creativity and diversity of examples on generated ideas still hold when we sampled examples using our scalable mechanisms. We presented participants with sets of example ideas that varied in creativity (as measured by a conventional method) and diversity (as measured by our idea map metric). Our results demonstrated that more creative examples led to more creative ideas being generated and that more diverse sets of examples led to more diverse sets of ideas being generated. However, we did not observe any impact of creative examples on the diversity of generated ideas or diversity of examples on the average creativity of generated ideas.

The work in this chapter made the following contributions:

- A crowd-powered method for automatically constructing an "idea map" that can be used to extract diverse sets of examples at scale.
- Validation of the generated "idea map'

- A study demonstrating that participants generated more diverse ideas when seeing diverse sets of examples generated using our idea map approach than when seeing randomly selected examples. The study also corroborated results from prior research by showing that people presented with particularly creative examples generated more creative ideas than those who saw set of random examples.

Together, these results can inform the design of systems to support large-scale collective ideation. Instead of leaving people to explore ideas of others haphazardly, future systems can help contributors to quickly find manageable sets of particularly creative and diverse ideas. Some existing systems already include mechanisms (such as voting mechanism used by OpenIDEO.com) for finding particularly creative ideas. We extend the state of the art by contributing a scalable crowd-powered approach that enables selection of sets of diverse ideas.

3.2 Scalable Mechanism for Identifying Diverse Sets of Ideas using an idea map

We need a way to construct sets of diverse ideas, and ideally also to systematically compare the relative diversity of pairs of sets. We only consider methods that incorporate human input, because fully automated methods currently tend to capture only superficial similarity [André et al., 2014]. Because we intend to use this measure in systems that support collaborative ideation in large groups, it also must scale to a large pool of ideas. We seek approaches that can be sustained by a large number of small contributions from non-experts. Moreover, it should be robust to between-rater differences in mental models and judgment calibration.

We chose to adapt an existing machine-learning–based method [Tamuz et al., 2011] that uses triplet similarity comparisons to place ideas in a two-dimensional map. The map is

Figure 3.1: An idea map generated by our system, showing emergent clusters of ideas around different themes and sentiments.

constructed such that ideas perceived by people to be similar to each other are placed close together, while ideas perceived to be very different are placed far apart. Figure 3.1 shows an example of such an idea map generated with our system.

To generate an idea map, we first present groups of three ideas to human judges and ask them to pick which of B or C is more similar to A. Compared to similarity rating query (how similar is A to B), this triplet based representation of relative similarity is less cognitively taxing to judges [Tamuz et al., 2011]. We use t-Distributed Stochastic Triplet Embedding (t-STE) [van der Maaten and Weinberger, 2012] to find an arrangement of ideas in a two-dimensional space (an "embedding") that is most consistent with the comparisons that people made. To minimize the number of comparisons that we ask people, we use an active learning heuristic [Tamuz et al., 2011] that estimates the expected gain in information about the position of an idea when comparing it to a particular pair of other ideas.

Informally, we expect the number of comparisons required to embed n ideas to be between $O(n)$ (scaling with the number of parameters of the model: 2 coordinates per idea) and $O(n \log n)$ (scaling with the number of comparisons required to find the closest existing idea for each new idea). For the most common ideas, even fewer comparisons should be required to determine that it is a common idea and thus unlikely to contribute much to the diversity of a set.

Because our idea maps are constructed such that the distances in the map reflect human perceptions of dissimilarity, we can use the maps to assess the uniqueness of an individual idea or the diversity of a subset of ideas. For example, we might define a unique idea as one that is far from other ideas. We use a simple metric of diversity: the diversity of a set of ideas is the mean distance between all pairs of those ideas.

3.3 Ideation Task and Seed Ideas

We collected an initial set of ideas from pilot studies. We used these ideas to validate our diversity measurement mechanism. We also used a subset of these as ideation examples for other participants in our main experiment.

Ideation Task The ideation task we chose for this study was to generate birthday messages for a greeting card for Mary, a female firefighter who is about to turn 50. The instruction for the task is included in Appendix A. We chose this task because it is short and simple, yet similar to the tasks of real creative professionals. Previous work in brainstorming and creativity has also used similar kinds of simple tasks [Guilford, 1967, Torrance, 1968, Smith et al., 1993, Marsh et al., 1996, Lewis et al., 2011]. Pilot experiments showed that the task was accessible to untrained participants, and that it elicited a wide variety of ideas of varying quality. We encouraged participants to generate lots of ideas within a 4-minute time limit and not to worry about the quality of the ideas. When they finished generating ideas, participants were asked to select a diverse set of up to 5 of their best ideas.

Participants For our pilot ideation study, we recruited 209 participants from 2 sources: our own social networks (63 participants) and MTurk (146 participants). For all MTurk studies in this paper, we limited recruitment to U.S. residents who had completed at least 1,000 HITs[1] with greater than 95% approval rate. A participant could do the task only once. MTurk participants were paid $1.50 for their participation, while uncompensated participants were given feedback on how the quantity and diversity of the ideas they generated compared to that of other participants.

[1] For triplet comparison tasks, only a minimum of 100 approved HITs were required.

3.4 Idea Map Elicitation and Validation

We then randomly selected 52 seed ideas from the 932 messages generated in the pilot studies from which to build an idea map.

3.4.1 Collecting Data to Build the Idea Map

We presented three birthday messages to each worker and asked him/her to pick which of the latter two ideas is more similar to/different from the first. We collected 2818 comparisons for 778 different triplets from 145 different people. We asked for multiple comparisons for the same triplets to enable subsequent analysis of inter-rater agreement. Many fewer comparisons would have been needed just to generate the idea map. The generated idea map is illustrated in Figure 3.1.

We then computed diversity scores for random subsets of the seed ideas. To illustrate, here are examples of idea sets to which our metric assigns *low* diversity scores:

- *"After 50 years your light is still burning strong", "We were worried you wouldn't be home on time, so we set your kitchen on fire.", "How many firefighters does it take to put out fifty candles?"*
- *"Wishing you a happy birthday!", "May the second 50 be as good as the first one!", "Happy Birthday!"*

While these idea sets get high diversity scores:

- *"Your cake is more lit up than a forest fire.", "Happy Birthday, Mary! 50 years is quite an accomplishment.", "Thank you for being there for us. Happy BD"*
- *"Have a fiery birthday bash!", "Time for Mary to start rolling down the hill!", "You have been one of a kind. Happy Birthday!"*

3.4.2 Validating the Idea Map

To validate the diversity ratings created by the idea map, we collected similarity ratings [Dow et al., 2010] for randomly chosen pairs of ideas in the example set. We then evaluated how well the measures of similarity captured in the idea map agreed with the perceptions of similarity provided by human raters.

We recruited 32 MTurk workers to rate similarity of pairs of messages on a scale of 1 (not at all similar) to 7 (very similar). Each rater rated about 30 pairs of messages. Each pair of messages was rated by three raters. We normalized (i.e., converted to z-scores) the ratings within each rater prior to aggregating the results. After excluding 4 workers whose answers to gold standard items indicated that they were not paying close attention to the task, we were left with 791 similarity ratings.

Krippendorff's alpha for the triplet comparison responses used to generate the idea map was 0.623 (nominal data) while the Krippendorff's alpha for the similarity ratings was 0.352 (interval data) indicating that comparison queries are, indeed, easier for participants to reach agreement on than rating queries.

Comparing mean human similarity ratings and our algorithm's diversity measure we found a significant correlation (Spearman correlation, $\rho = -0.5284$, $p < .0001$). Note that our measure captured diversity while the participants were asked to assess similarity, so the negative correlation coefficient is the desirable outcome.

Krippendorff's alpha between mean z-scored similarity ratings (standardized, sign of similarity inverted) and the diversity measure generated by our algorithm was 0.55. This is a high level of agreement considering that human raters agreed with each other only with alpha = 0.35.

3.5 Main Experiment

We designed our main experiment to explore the possibility of having a large scale collaborative idea generation system where judiciously chosen ideas from previous contributors are used as ideation examples for newcomers. Namely, we want to look at the effects of creativity and the diversity of ideation examples–algorithmically sampled from a pool of ideas based on intended intervention–on the creativity and diversity of ideas produced by later participants.

3.5.1 Tasks

We used the same ideation task as in the pilot study: generate birthday messages for Mary, a firefighter who is turning 50. With 20% probability, participants were asked to perform exactly the same task as in the pilot study, while the others were presented with an intervention: At the beginning of the ideation task, they were shown a set of 3 example ideas (which remained visible throughout the idea generation phase).

Interventions We used the same set of 52 ideas generated in the pilot study as possible ideation examples. We varied the individual creativity of the ideation examples as well as the diversity of the sets of examples to investigate how these manipulations impacted individual idea generation.

Two trained coders from our research team independently rated the creativity of each birthday message on the scale from 1 (not creative) to 3 (very creative). We marked as "creative" the eleven messages that received scores of at least 2 from both coders. To illustrate, some of the most creative messages were: *"We were worried you wouldn't be home on time, so we set your kitchen on fire."* and *"How many firefighters does it take to put out a birthday cake?"*, while some of the least creative were: *"Hey Mary, It's Your Birthday, Happy Birthday!"*

and *“Love and Happiness to Mary, one of the best!”*

Half of the participants who were presented with ideation examples saw messages sampled only from the pool of 11 creative messages (*Creative examples only* condition), while those in the other group saw ideation examples drawn uniformly from the entire pool of 52 ideas (*All examples* condition).

To investigate the impact of diversity of ideation examples on individual ideation outcomes, we used the diversity metric introduced in the previous section to assess the diversity of each randomly generated set of ideation examples presented in either of the creativity conditions. The mean diversity score in the All examples condition (M=9.85) was higher than in the Creative examples only condition (M=8.71), but the difference was small (Cohen's $d = 0.36$). The variances of diversity scores in the two creativity conditions were similar. There is no statistically significant difference of diversity of examples between the two conditions (t(116)=1.84, n.s.)

3.5.2 Procedure

As in the pilot experiment, each participants had 4 minutes to generate as many messages as they could, and they selected up to 5 as a diverse set of their best ideas.

To measure whether participants paid attention to the given examples (and thus could have been influenced by our manipulation), at the end of the experiment we showed them five ideation examples and asked them to select the ideas they saw during the ideation tasks. Three of the five ideas were the ideas that had been shown at the previous stage while the other two were distractors.

3.5.3 Design And Analysis

For the primary analysis we used a 2×2 full factorial between-subject design with the following factors and levels:

- *Creativity of ideation examples* {All examples, Creative examples only}
- *Diversity of ideation examples* (modeled as a continuous variable).

Our measures were:

- *Creativity of generated ideas* assessed by expert raters.

  Five experts from oDesk rated creativity of generated ideas. All experts were professional writers or editors. Each expert rated at least 300 messages on a scale from 1 (not at all creative) to 7 (very creative). Each message was rated by three experts. Our creativity measure is the average of each expert's normalized rating for each message.

- *Diversity of generated ideas* assessed by MTurk workers.

  We chose to use an established measure of diversity for our outcomes: as in the validation experiment, we used average pairwise similarity [Dow et al., 2010]. We randomly selected 15 participants from the baseline condition, 30 participants from the All examples conditions and 30 participants from the Creative examples only condition. For this measure, we only included participants who generated more than one idea. We only analyzed messages that participants included in their diverse sets of best messages. For each participant, we asked 3 workers to rate the similarity of each pair of generated ideas. As before, we converted worker ratings into z-scores prior to analysis. We flipped the sign of z-scored similarity ratings to derive diversity scores.

For each measure, we conducted an analysis of covariance including both factors and their interaction.

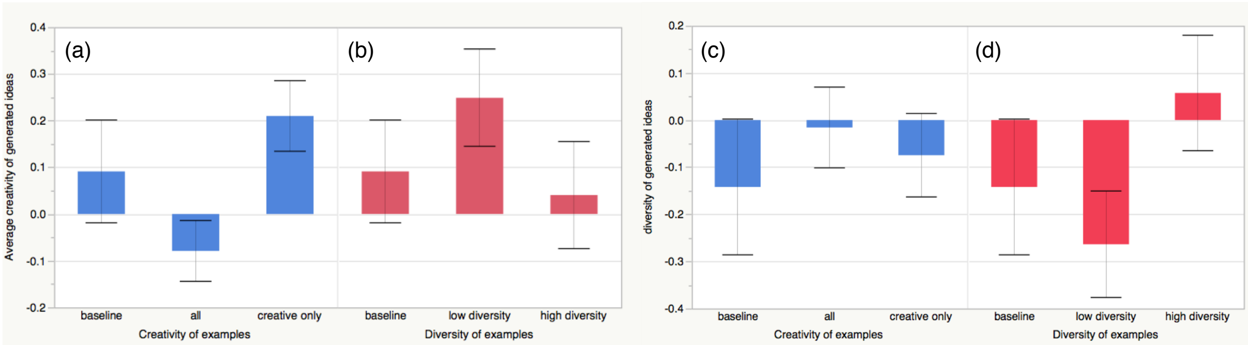


Figure 3.2: (a) Participants in the Creative Only condition generated more creative ideas than participants in the All condition. (b) There is no difference in average creativity of generated ideas across groups seeing different levels of diversity. (c) There is no difference in the diversity of generated ideas across groups seeing different levels of creativity. (d) Participants who saw examples with high diversity generated more diverse sets of ideas than those who saw examples with low diversity.

We also compared our interventions to the baseline condition. For the *Creativity of ideation examples* factor, we conducted an analysis of variance with one factor with three levels: baseline, All examples and Creative examples only. For the *Diversity of ideation examples* factor, we first created two discrete diversity conditions: Low diversity (which included the ideas generated by participants who saw the 25% least diverse sets of ideation examples) and High diversity. We then conducted an analysis of variance with one factor with three levels: baseline, Low diversity and High diversity.

3.5.4 Participants

We recruited 138 participants via MTurk to generate the ideas under the same recruitment limitation as in the pilot experiment.

Three participants did not complete the task and were excluded from further analysis. There were 27 participants in the baseline condition, 49 in the All examples condition and 59 in the Creative only examples condition.

Adjustments of Data

We filtered out participants who did not pay attention to the examples — those who answered correctly fewer than four out of five questions when asked which ideation examples they saw while ideating. After the exclusion, there were 27 participants in the baseline condition, 48 in the All examples condition and 52 in the Creative only examples conditions.

3.5.5 Results

127 participants generated 723 ideas and selected 564 ideas to be their best ideas. We only analyzed the 564 self-selected ideas. For the similarity assessment, 52 workers generated 1,581 ratings.

Creativity of generated ideas

We observed a significant main effect of creativity of ideation examples on the mean creativity of generated ideas, F(1,96)=6.95,$p = 0.0098$. Participants who were presented with Creative only ideation examples produced ideas that received higher mean creativity scores (M=0.21) than participants who were presented with randomly selected ideation examples (M=-0.079). However, the example diversity had no significant effect on the creativity of generated ideas (F(1,96)=1.13, n.s.).

In a three-way comparison between the baseline condition and the two creativity conditions (Figure 3.2a), we observed a significant main effect of condition on creativity of generated ideas (F(2,124)=3.91, $p = 0.0227$). Participants who were presented with Creative only ideation examples had higher scores (M=0.21) than people in the baseline condition (M=0.0912), while participants who were presented with random examples had lower scores (M=-0.079) than participants in the baseline condition. A post hoc Tukey HSD test showed that neither of these two pairwise differences was significant, however. The significant difference responsible for the main effect was between the Creative only and All examples conditions.

In a comparison of the baseline condition to participants who saw the High diversity and Low diversity example sets (Figure 3.2b), participants who saw diverse examples generated ideas with slightly lower creativity scores (M=0.0406) than participants in the baseline condition (M=0.0912) while those who saw least diverse examples had higher creativity scores (M=0.249) than participants in the baseline condition. However, this effect was not significant (F(2,75) = 0.99, n.s.).

Diversity of generated ideas

We observed a significant main effect of the example diversity on the mean diversity of generated examples (F(1,56)=2.26, $p = 0.028$) with diversity of generated ideas increasing with the increase in the diversity of examples. However, we observed no significant effect of diversity of examples on the creativity of generated ideas (F(1,56)=3.33, n.s.)

In a three-way comparison between the baseline condition and the two creativity conditions (Figure 3.2c), we observe no significant effect of the creativity of examples on the diversity of generated ideas (F(2,72) = 0.34, n.s.). The three way comparison including the baseline condition and the participants who saw the High diversity and the Low diversity examples (Figure 3.2d) also produced no significant effect (F(2,41) = 1.56, n.s.)

3.5.6 Additional Analyses

The results so far show that people generate creative ideas when they see creative examples and that they generate a diverse set of ideas when they see a set of diverse examples. But are people genuinely motivated and inspired by the examples (as suggested by [Marsh et al., 1996, Paulus and Dzindolet, 1993, Leggett Dugosh and Paulus, 2005]), or do they simply produce ideas that closely imitate the examples?

To answer this question, we measured how similar the generated ideas were to provided examples. Specifically, we asked 130 MTurk workers to rate similarity of generated messages to the examples using the same procedure as in the validation experiment. For each generated idea, we found the closest example out of the three examples that the participant saw. High similarity to the closest example indicates high degree of fixation. Averaging similarity to the closest example for each of the participant's ideas, we get a measure of how similar the ideas this participant generated were to provided examples.

For the baseline condition (where no examples were given), we measured *self-fixation* [Nijstad and Stroebe, 2006] instead: that is, we measured how similar each new idea was to the closest of the ideas the participant had already generated. While not directly comparable to the fixation induced by externally-provided examples, this measure provides an informative baseline for evaluating how much external examples influenced each participant's ideas.

Rather than fixating participants, we found that good example sets actually did the opposite. Participants in the 'Creative only' condition generated ideas that were rated less similar to the examples (M=0.43) than the participants in the 'All examples' condition (M=0.65, t(98)=2.49, p=0.0143) (Figure 3.3a). Likewise, participants in the 'High diversity' condition generated ideas with lower similarity to most similar example (M=0.41) than the participants in the 'Low diversity' condition (M=0.68, t(46)=2.30, p=0.0260) (Figure 3.3b). In both interventions, the similarity to examples was lower than the self-fixation observed in the baseline condition (M=0.79).

We also manually inspected the ideas generated by 20 participants randomly sampled from all but the baseline condition and we compared the ideas they generated to the examples they were shown. The results suggest that participants often generated ideas seemingly entirely unrelated to the examples or added a new spin on an example (e.g., a participant who saw *"How many firefighters does it take to put out fifty candles?"* generated *"Get ready to call the fire department, we are about to light the 50 candles!!"*). They sometimes tried to combine ideas from more than one examples (e.g., a participant who saw *"We were worried* *you wouldn't be home on time, so we set your kitchen on fire."* and *"Remember, blow out the candles on your cake, don't use the hydrant!"* generated *"Mary! That's a lot of candles! If the place catches on fire, at least we won't have to call anyone!"*). There were cases of surface feature borrowing (e.g., a participant who saw *"Mary you could rescue me any day!"* generated *"mary you could put out fires for me any day"*), but such cases were rare.

These additional analyses suggest that there is no evidence that presenting participants with examples stifled their creativity. Instead, the results provide additional evidence that presenting people with particularly creative or particularly diverse ideas may help: those participants generated ideas that were more original (i.e., less similar to the examples) than the participants who saw more mundane examples.

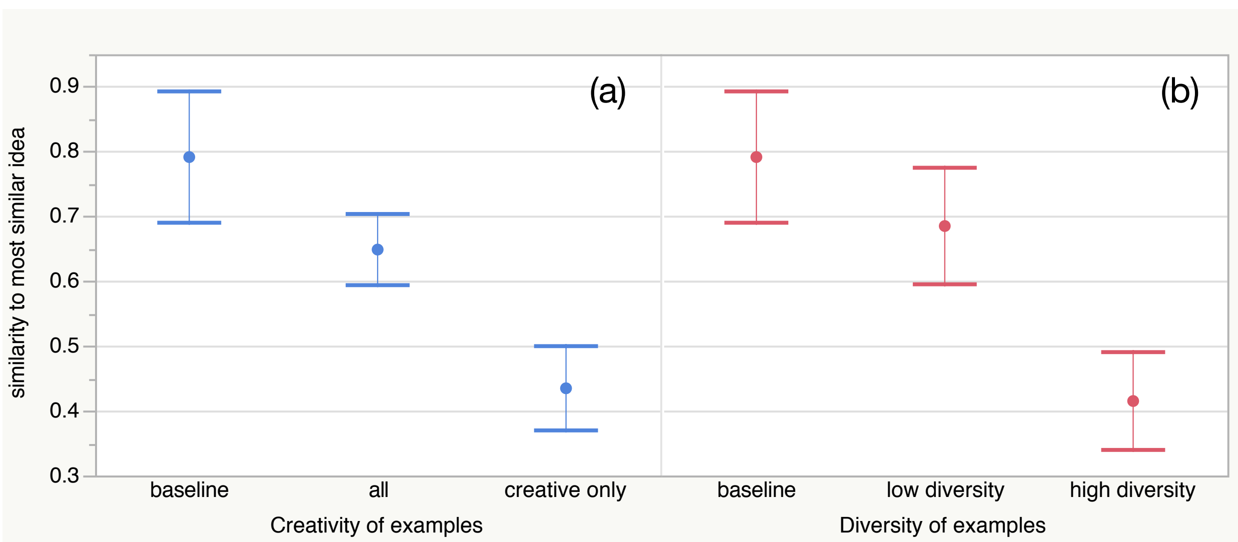


Figure 3.3: (a) Participants in the Creative only condition were less fixated than those in the baseline and the All examples condition. (b) Participants who saw a set of diverse examples were less fixated those in the baseline condition and those who saw a set of examples with low diversity.

3.6 Discussion

Our studies demonstrate that we can select sets of diverse examples using a scalable method, and that people presented with the examples so selected generate more diverse ideas than those presented with random examples. Similarly, seeing examples of ideas that others deemed as particularly creative improves the creativity of generated ideas compared to seeing randomly selected examples.

Neither intervention resulted in ideation outcomes that were statistically different from not showing any examples at all, but the trends were illuminating: participants who saw creative ideation examples produced more creative ideas than those who saw no examples at all, but participants who saw randomly selected examples produced the least creative ideas. We observed a similar trend for diversity: participants who saw the 25% most diverse sets of examples produced more diverse ideas than participants in the baseline condition, while participants who saw the 25% least diverse sets of examples produced the least diverse sets of ideas of all participants.

Two possible explanations of the results arise. One explanation is that people get inspired by example ideas and incorporate these examples in their own idea generation. This explanation implies that we can guide how a community explores the space of possible ideas by exposing people to ideas in particular areas of interest. Another explanation involves social influence. People might infer the desirable properties of a set of ideas from the example set that they saw. Here, an example set provides information about the performance of others, encouraging participants to match the properties of their own ideas to example sets [Paulus and Dzindolet, 1993, Leggett Dugosh and Paulus, 2005]. While the two explanations involve very different mechanisms, they both support the value of presenting users with sets of creative and diverse examples. In order to understand which is the more likely cause, we need to conduct further investigation. For example, a future study can ask participants about the desirable properties of a set of generated ideas and how they use examples to infer whether they just try to match the properties of an example set or whether they actually incorporate the content of the examples into their own ideation.

Despite contrasting explanations, our results demonstrate the feasibility and value of using scalable crowd-powered mechanisms to improve large-scale online collaborative ideation platforms: instead of leaving contributors to manually browse through hundreds or thousands of previously generated ideas, these systems can help contributors by selecting manageable sets of particularly creative and diverse ideas.

One limitation of our work is that we have only studied the effect of showing people the raw ideas that others generated. Alternative interventions include presenting categories or schemas (as in [Yu et al., 2014b]), or giving specific instructions about what kind of idea to generate.

Another limitation of our work is timing: the best time to present people with inspirational examples might be when they run out of their own ideas, not right at the beginning of the ideation process.

Finally, we suspect that the 4-minute time limit might prevent some participants from putting in enough cognitive effort to process examples deeply enough to benefit from them.

3.7 Conclusion

One challenge in designing large-scale collaborative online ideation platforms is how to leverage the ideas generated by others to effectively inspire future (or returning) contributors. As prior research suggests and as our results corroborate, showing people random examples of prior ideas has little positive impact on what new ideas people generate. However, prior research suggests that presenting people with sets of particularly creative or particularly diverse ideas is likely to improve the creativity and diversity of generated ideas.

These prior findings were not easy to act on: while there exist scalable crowd-powered methods for identifying the most creative ideas among thousands, the same is not true for finding sets of *diverse* ideas. In this paper, we contribute a scalable method for evaluating *diversity* of sets of examples by using simple similarity comparisons from non-expert contributors (members of the ideation community or an external crowd) to create an *idea map*. An idea map is a two-dimensional embedding of the ideas such that the pairwise distances between ideas on a map correspond to human perception of dissimilarity. Idea maps make it possible to sample sets of ideas of varying levels of diversity by picking ideas that are close to or far from each other.

The results of our study show that this method is indeed effective: participants who saw sets of diverse examples generated using our method produced more diverse ideas than participants who saw randomly selected examples. Our study also corroborates previous findings that showing people examples that others consider particularly creative results in more creative ideas than showing random ideas.

The goal of this chapter was to inform the design of future systems for supporting collaborative ideation at large scale. Our scalable method for assessing diversity, together with existing creativity metrics [Salganik and Levy, 2012, Yu and Nickerson, 2011, Tanaka et al., 2011, Xu and Bailey, 2012], can enable creativity support systems to adjust which examples are shown to contributors and thus, as we have shown, modulate the quality and diversity of the ideas that they contribute. These methods, thanks to their lightweight nature, can be either outsourced to external micro-task market or embedded in the ideation workflow where contributors provide information about the example ideas for succeeding contributors.

Chapter 4

IdeaHound: Integrated-crowdsourcing for Creativity Enhancing Interventions

This chapter has adapted, updated, and rewritten content from a paper at UIST 2016 [Siangliulue et al., 2016]. All uses of “we”, “our”, and “us” in this chapter refer to coauthors of the aforementioned paper.

Prior work on creativity support tools demonstrates how a computational semantic model of a solution space can enable interventions that substantially improve the number, quality and diversity of ideas. However, automated semantic modeling often falls short when people contribute short text snippets or sketches. Innovation platforms can employ humans to provide semantic judgments to construct a semantic model, but this relies on external workers completing a large number of tedious micro tasks. This requirement threatens both accuracy (external workers may lack expertise and context to make accurate semantic judgments) and scalability (external workers are costly).

In this chapter, we introduce IdeaHound, an ideation system that seamlessly integrates the task of defining semantic relationships among ideas into the primary task of idea generation. The system combines implicit human actions with machine learning to create a computational semantic model of the emerging solution space. The integrated nature of these judgments allows IdeaHound to leverage the expertise and efforts of participants who are already motivated to contribute to idea generation, overcoming the issues of scalability inherent to existing approaches. Our results show that participants were equally willing to use (and just as productive using) IdeaHound compared to a conventional platform that did not require organizing ideas. Our integrated crowdsourcing approach also creates a more accurate semantic model than an existing crowdsourced approach (performed by external crowds). We demonstrate how this model enables helpful creative interventions: providing diverse inspirational examples, providing similar ideas for a given idea and providing a visual overview of the solution space.

4.1 Motivation and Contributions

Large creative online communities will transform the way our society innovates. Existing communities, like OpenIDEO (openideo.com), where people propose solutions to social problems, and platforms, like coUrbanize (courbanize.com), where cities gather ideas from their citizens, attract large numbers of users, many of whom contribute ideas or designs. The promise of these online communities is that participants will benefit from exposure to ideas of others and, thus inspired, will generate better ideas than they would have otherwise. In practice, however, crowd innovation challenges result in large quantities of simple, mundane and repetitive ideas [Bjelland and Wood, 2008, Klein and Garcia, 2015, Riedl et al., 2010]. Consequently, many organizations have come to see crowd innovation platforms more as marketing gimmicks that energize their customers or constituents, rather than real sources of innovation. Meanwhile, numerous creativity-enhancing interventions targeted at individuals

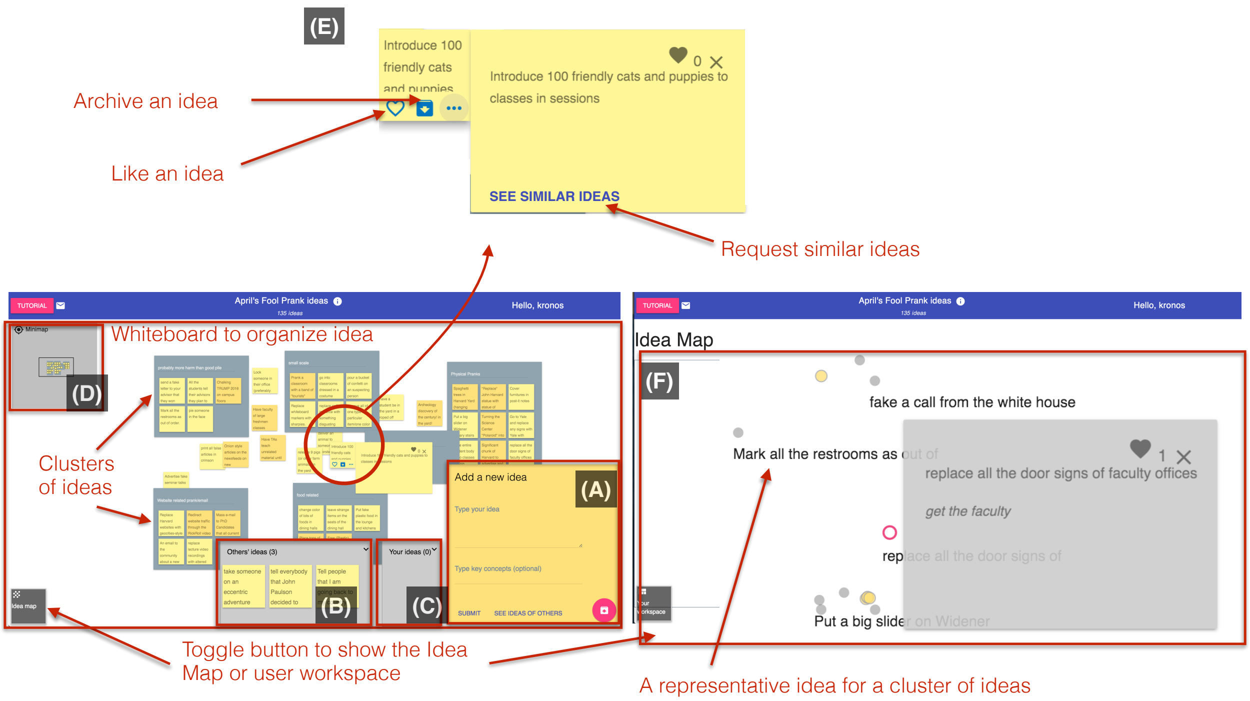


Figure 4.1: IdeaHound interface. (A) A box where users can type and submit their ideas; (B) When users request ideas from others, they appear on the Others' Ideas pane, (C) When users submit an idea, it first appears on the Your Ideas pane; Users can move ideas from (B) and (C) to organize on the whiteboard area. When they place ideas close to each other, a cluster will form around the ideas. (D) A minimap of the workspace. Users can pan and zoom the whiteboard or control the zoom from the minimap view. (E) When they hover over an idea, a control panel allows users to like the idea, remove the idea from the workspace, or open up a Details pane for that idea. On the Details pane, users can click "See similar ideas" to request ideas of others that are similar to that idea. (F) The idea map visualization is a 2D map that gives an overview of the solution space. Each dot represents an idea. The user's own ideas are in orange while the ideas from others are in yellow. A label for each cluster of ideas on the idea map visualization shows a sampled idea from that cluster.

and small groups exist, and many of these interventions have been demonstrated to measurably improve the creative outcomes. How might we build on these successes to improve the quality and diversity of ideas contributed on large scale collaborative ideation platforms?

Many creativity-enhancing interventions leverage corpora of relevant design examples *and* a computational insight into the structure of the solution space revealed by those examples. For example, Design Gallery for 3D modeling [Marks et al., 1997] and other similar systems [Talton et al., 2009, Lee et al., 2010] help users gain a quick intuition of the solution space and facilitate recombination of disparate ideas [Nickerson et al., 2008] by showing them multiple *diverse* alternatives. ReflectionSpace [Sharmin and Bailey, 2013] and Freed [Mendels et al., 2011] support reflection in the design process by presenting users' designs in the context of other related artifacts. Adaptive Ideas web design tool [Lee et al., 2010] and DesignScape [O'Donovan et al., 2015] promote broad exploration of the solution space during the divergent phase of idea generation by showing a diverse set of examples and design alternatives. They also support refinement by allowing users to explore sets of closely related ideas, all of which pursue the same general approach, but in subtly different ways [Lee et al., 2010, O'Donovan et al., 2015].

All of these systems leveraged some computational representation that made it possible to tell which ideas were similar to each other and which were different. They either leveraged the fact that the design space was parameterized to begin with (e.g., 3D models in [Marks et al., 1997]) or they used some mechanism to automatically compute descriptive features of the artifacts (e.g., [Talton et al., 2009, Lee et al., 2010]). On existing large scale collaborative ideation platforms, people tend to communicate their initial ideas in the form of short text snippets or sketches. Thus, no *a priori* parametrization of the solution space is available. Furthermore, feature discovery mechanisms such as probabilistic topic modeling [**?**] do not perform well with such representations [Chang et al., 2009, Chuang et al., 2012].

Crowd-powered systems offer a possible solution: by judiciously combining human judgement and machine learning, it is possible to discover useful structure in collections of arbitrary artifacts [André et al., 2014, Chilton et al., 2014, Siangliulue et al., 2015a, Tamuz et al., 2011]. However, existing crowd-powered approaches have a crucial limitation on their applicability to large-scale collaborative ideation platforms: they depend on people completing a large number of tedious and repetitive micro-tasks. This requirement means platforms that seek to leverage such approaches must employ large numbers of external workers (e.g., from online labor markets such as Amazon Mechanical Turk or UpWork). This is not a desirable approach, for two main reasons. First, employing large numbers of workers is expensive, which limits the ability of these systems to scale to very large innovation platforms. Secondly, even if cost was not a concern, many online creative communities assume some amount of shared knowledge (e.g., local knowledge among contributors to a municipal participatory budgeting platform), which would not be available to workers hired outside the community. Thus, the human judgements on semantic relationships among ideas should come from the creative community itself.

This chapter's first study suggests that it is infeasible to expect unpaid, intrinsically-motivated participants to complete a secondary task of judging ideas of others in addition to the primary task of generating ideas. We recruited unpaid, intrinsically motivated participants to generate ideas and we then asked them to evaluate ideas generated by other members of the community (rate similarity between ideas and idea quality). When we required participants to complete these evaluation tasks, they found the tasks to be tedious and repetitive; when the completion of the tasks was voluntary, participants did not complete enough of those tasks to inform the creation of a reliable computational model.

In response to this challenge, we designed IdeaHound, a self-sustainable system for supporting creative ideation at scale. A crucial, novel component of IdeaHound is an *integrated crowdsourcing* approach that seamlessly integrates the potentially tedious secondary task of analyzing semantic relationships among ideas with the more intrinsically-motivated primary task of idea generation. Our integrated approach leverages the insight that people naturally tend to spatially organize their inspirational material (including their own ideas) such that ideas and inspirations that share something in common are grouped together. IdeaHound thus presents users with a prominent affordance for spatially organizing their own ideas and ideas of others. IdeaHound continuously monitors the evolving spatial organizations created by all members of the community and creates a global model capturing relative similarities and differences among ideas. This model can help the community accomplish tasks both during idea generation (e.g., finding inspirations and gaining overview of solution space) and after idea generation (e.g., organizing ideas and selecting ideas). Figure 4.1 illustrates the main features of IdeaHound.

Our empirical studies demonstrate the viability of this approach. In Studies 2 and 3, participants implicitly defined semantic relationships among ideas by spatially organizing their own ideas and those of their peers while they were generating novel ideas. The results of these studies demonstrate that, even though participants were not explicitly asked to spatially organize ideas, they naturally did so frequently and thoughtfully enough to create an accurate computational model of the semantic relationships among ideas. The resulting model agreed with standard (and more expensive) human judgements more closely than a computational model created using a conventional outsourcing approach [Siangliulue et al., 2015a], where a separate crowd (of equally qualified participants) completed stand-alone semantic judgment tasks. Further, participants generated as many ideas (despite doing the extra work arranging ideas) and were as satisfied with the *integrated crowdsourcing* interface as they were with an equivalent conventional interface that required no additional work besides submitting their own ideas and browsing the ideas of others.

We demonstrate how the resulting semantic model can be used to enable three creativity-enhancing interventions in IdeaHound: sampling diverse inspirational examples, exploring similar ideas, and providing a visual overview of the emerging solution space. In Study 4, we conducted a preliminary end-to-end evaluation of IdeaHound. The results show that people found the suggested diverse sets of ideas helpful for their idea generation. They also found that the map visualization provided them with a quick and useful overview of the evolving solution space.

This chapter makes the following contributions:

1. A crowdsourcing approach that integrates the potentially tedious task of evaluating

creative ideas with the more exciting task of idea generation, such that contributors, who are intrinsically motivated only to contribute to idea generation, perform both tasks.

2. An end-to-end system, called IdeaHound, which uses crowd-contributed spatial arrangements of ideas to construct a robust model of semantic relationship among ideas. IdeaHound uses this model to enable three creativity-enhancing interventions: sampling diverse inspirational examples, exploring similar ideas, and providing a visual overview of the emerging solution space. While similar interventions were previously used to enhance the performance of individuals and small groups, IdeaHound makes it possible to support creative communities of hundreds or thousands of contributors.

3. Empirical studies that demonstrate the need for and the viability of an integrated crowdsourcing approach for supporting enhanced collective ideation at scale.

4.2 Design Goals

The end goal of this work is to improve large-scale collaboration with creativity-enhancing interventions, such as providing diverse inspirational examples, enabling exploration of similar ideas (for iteration), and providing a real-time "map" or overview of the solution space. As we have seen in the review of prior work, these interventions depend on having access to a semantic model that captures the structure of the solution space (e.g., how solutions relate to each other). However, none of the existing solutions for constructing such models are adequate: the completely automated approaches are unlikely to work well with short text snippets and sketches, while the crowd-powered solutions depend on large numbers of external workers completing many tedious/repetitive semantic judgment tasks.

Therefore, the technical focus of this work is to create an approach for semantic modeling of solution spaces that meets two main requirements:

1. *Nearly Real-time.* The approach should be able to provide a nearly real-time model of the solution space.
2. *Self-sustainable.* The approach should not depend primarily on external labor.

Our general approach is to combine methods from crowdsourcing and machine learning research. Specifically, similarly to [Siangliulue et al., 2015a], we rely on a modest number of human judgements regarding relative similarities of pairs of ideas and we then use machine learning techniques to efficiently combine those human judgements into a consistent and comprehensive model of the emerging solution space. Unlike the prior work, however, we seek to engage the members of the creative community themselves in the process of constructing the semantic model instead of outsourcing the task to external crowds. In the following sections, we describe the rationale, design, technical details, and evaluation of our approach. We also demonstrate how this approach allowed us to build IdeaHound, an end-to-end self-sustainable system that enables three creativity interventions for enhancing collective ideation at scale.

4.3 Study 1: Separate Tasks to Collect Semantic Relationships Among Ideas

A straightforward approach to solicit the necessary human judgements of semantic relationships among ideas is to explicitly ask the members of the community to contribute these judgements. We tested this approach in a study conducted on the LabintheWild.org platform [Reinecke and Gajos, 2015], which attracts intrinsically motivated, unpaid online participants who take part in studies in return for informative feedback on their performance. We recruited 2,061 participants to generate ideas for birthday messages for a 50-year-old female firefighter. The study had four parts: 1) participants generated as many ideas as they could in 4 minutes, 2) after they finished generating ideas, they were asked to provide a small set (5–10) of human judgments of semantic similarity between ideas (using the same mechanism as [Siangliulue et al., 2015a]), 3) they were presented with a results page, and 4) they ranked ideas of others based on their quality. The last part of the experiment was optional and participants could skip this part at any time.

141 of the 2,061 participants who finished generating ideas in part 1 dropped out before completing the semantic judgments in part 2. Further, fewer than half of the participants (743 out of the remaining 1,920) finished the optional ranking task in part 3. Some participants noted in their post-study open-ended comments that the semantic judgement and ranking tasks were repetitive, unappealing and took too much time. One participant almost gave up on the semantic judgment task because it was "boring and cruel". This suggests that when given a choice to optionally complete these extra human judgment tasks, few participants on these platforms will choose to do so.

4.4 Integrated Crowdsourcing of Creative Ideas and Semantic Relationships

Instead of asking users to provide insights into a solution space by doing tedious tasks that detract from generating ideas, we sought to design an interaction that seamlessly integrated subjective judgement tasks with idea generation. Figure 4.1 shows the main interface for the final prototype.

In designing this solution, we drew inspiration from several existing systems, which require diverse kinds of work to be accomplished, but whose users are intrinsically motivated to do only a subset of those tasks. For example, Duolingo integrated the potentially tedious secondary task of translating real world text with the intrinsically valuable primary activity of learning a new language. In the CROWDY system [Weir et al., 2015], people who want to learn specific skills (e.g., web programming) improve video tutorials for future learners (the secondary task) as a byproduct of learning from those tutorials (the primary task). The users of the American Sign Language (ASL) flashcard quiz [Bragg et al., 2015] improve the feature-based indexing of the signs for the new ASL dictionary (the secondary task) as a byproduct of practicing the signs (the primary task). In another system [Komarov and Gajos, 2014], students generate formative feedback on each other's assignments (the secondary task) as a byproduct of studying for an exam (the primary task). While all of these prior systems leveraged the users' desire to learn, we believe the approach of integrating a valuable but potentially tedious secondary task into an intrinsically motivating task generalizes to other settings where users have different intrinsic motivations.

In the rest of this section, we describe the iterative development of our integrated crowdsourcing approach through a series of formative prototypes.

4.4.1 Initial Design: Continuous Spatial Arrangement

We based our design on the insight that people naturally spatially organize their inspirational material. The key feature of our design is a whiteboard space where users can arrange their own ideas or ideas of others. Because the spaces where users' own ideas and the inspirational examples first appear are very small, users naturally tend to drag ideas (their own and those of others) onto the canvas and organize them spatially. Because the whiteboard naturally affords continuous spatial arrangements (placing ideas close or far from each other), we initially built on the SpAM approach of collecting similarity information from people's spatial arrangements [Goldstone, 1994, Hout et al., 2013]. From each spatial arrangement generated by a user, our system extracted similarity scores from relative distances between pairs of ideas. Then the system aggregated these implicit similarity judgements from users' whiteboards using multidimensional scaling (MDS) algorithm to generate an aggregate semantic model.

Our pilot study with this version of the prototype showed that people naturally organized ideas spatially without instructions to do so, but not in the way that the system was designed for: Instead of organizing ideas such that physical distances among them would represent the degree of dissimilarity, as SpAM assumes, people tended to aggregate ideas into discrete clusters. Open-ended comments from participants revealed that such discrete clustering (rather than continuous spatial arrangement) gave them a better sense of the emerging themes and provided a more readable "big picture" of the possible approaches to the creative challenge at hand.

4.4.2 Revised Design: Explicit Clustering

In our second design we made cluster-forming actions explicit. Whenever a user brings two ideas into close proximity, an outline is drawn around both ideas to indicate that they are now grouped into a cluster. Cluster management is fluid yet explicit: when a user brings an idea close to an existing cluster, the cluster automatically expands; when the user drags an idea away from a cluster, the idea is removed from the cluster. Although it is possible for an idea to belong multiple clusters, a user can put it in only one cluster to keep the interaction simple.

According to feedback from our pilot studies, this approach was intuitive and matched users' expectations well. However, they reported that they sometimes forgot what concept they had intended to capture with each cluster. This was particularly frustrating to the users when ideation was performed over the course of several days: when they returned to the task after a day's break, they had a difficult time remembering the organizational structure they had been working to create.

Also, when we analyzed clusters created as part of several studies, it became clear that not all clusters were used to capture semantic similarity. Instead, some clusters were used to store "other" ideas or user's own ideas regardless of their semantics. This was problematic because it created a mismatch between the actual semantics of some of the clusters and the assumptions made by our algorithm.

4.4.3 Final Design: Explicit Clustering with Labels

To address these two issues, in our final prototype we introduced a clear affordance to add optional textual labels to clusters. This design turned out to be very effective. Not only did it help participants remember better what each cluster was intended to capture, it also substantially reduced the number of clusters that did not capture semantic similarity. Thus, this design choice simultaneously made the spatial organization capability more useful to the users and made the user-generated clusters a more valuable source of data for the machine learning algorithm.

4.4.4 Computational Model

As illustrated in Figure 4.2, to compute a global computational representation of how similar or different the collected ideas are from each other, our system initially constructs a similarity matrix from clusters across the users. Here, the similarity between two ideas is the empirical probability that the two ideas will be in the same cluster if they are both placed on the same whiteboard. This similarity matrix is sparse, however: not all pairs of ideas appear on the same whiteboards, so not all pairwise similarities are estimated. Therefore, the system computes an approximate idea similarity matrix (but one that estimates all pairwise similarities) using the t-Distributed Stochastic Neighbor Embedding (t-SNE) algorithm [Van Der Maaten, 2014]. Following [Siangliulue et al., 2015a], we refer to this embedding as an *idea map*. This embedding provides an approximate estimate of similarities among all pairs of ideas, for which at least some similarity data are available.

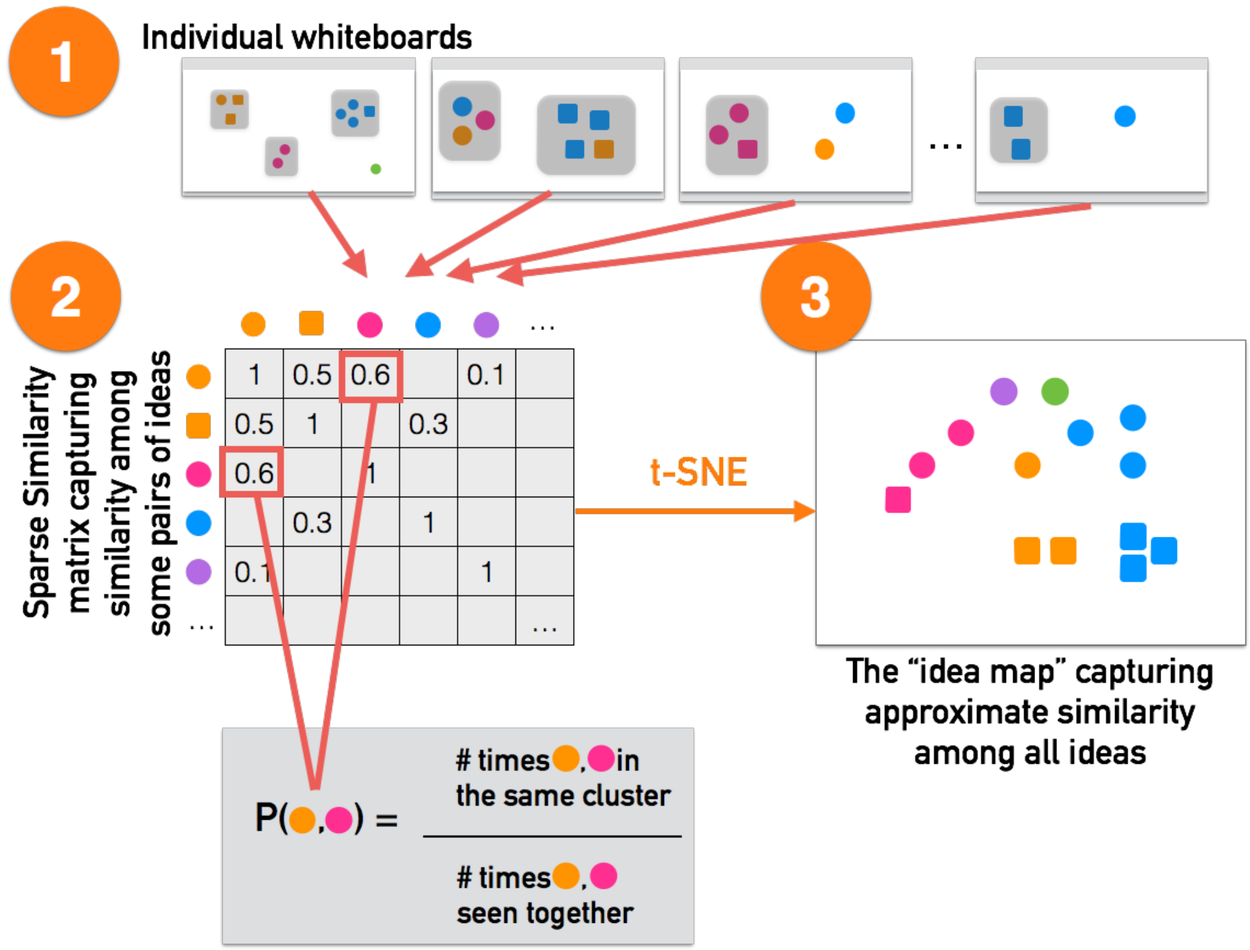


Figure 4.2: Computational model generation process. The system 1) aggregates grouping information from all users' whiteboard organization, 2) constructs a sparse similarity matrix from aggregated grouping, and 3) generates an "idea map" that puts similar ideas closer to each other and keeps dissimilar ideas far from each other according to similarity matrix in 2).

4.5 Evaluation of the Technical Approach

The central goal of our approach is that the potentially tedious secondary task of organizing ideas be integrated seamlessly into the intrinsically motivating primary task of idea generation. We evaluated our approach in two ways. In Study 2, we evaluated the experience and creative output of the users who used the system with our integrated crowdsourcing approach, compared to users who used a conventional interface. In Study 3, we evaluated the accuracy of the semantic model created using our integrated approach by comparing it to a model generated using a previously-validated method [Siangliulue et al., 2015a] that relies on outsourced crowd workers.

4.5.1 Study 2: User Experience and Creative Output with the Integrated Crowdsourcing Approach

In this study, we compared the experience and creative output of the users who used the integrated crowdsourcing approach, to users who used a conventional interface. We hypothesized that there would be no difference in experience and creative output between those who ideated with the integrated crowdsourcing interface and those who used a conventional interface.

Design

We used a between-subjects design with one factor with two conditions:

- *Integrated*: Participants used the integrated crowdsourcing interface like the one shown in Figure 4.1 to generate ideas. They could request to see ideas of others by clicking on the "SEE IDEAS OF OTHERS" button (Box A of Figure 4.1). The system then presented a set of up to three ideas. From ideas for which the system had information, the system sampled the first and the second idea. The first idea was selected randomly and the second idea was the idea that was predicted to be the most different from the first idea; the third idea was sampled randomly from ideas for which the system had no information. If there were no more unseen ideas, the system asked the user to request ideas again later. Participants could organize their own ideas and ideas of others together on the whiteboard. Unlike the interface in Figure 4.1, the participants could not request to see similar ideas to an idea or look at an idea map visualization.

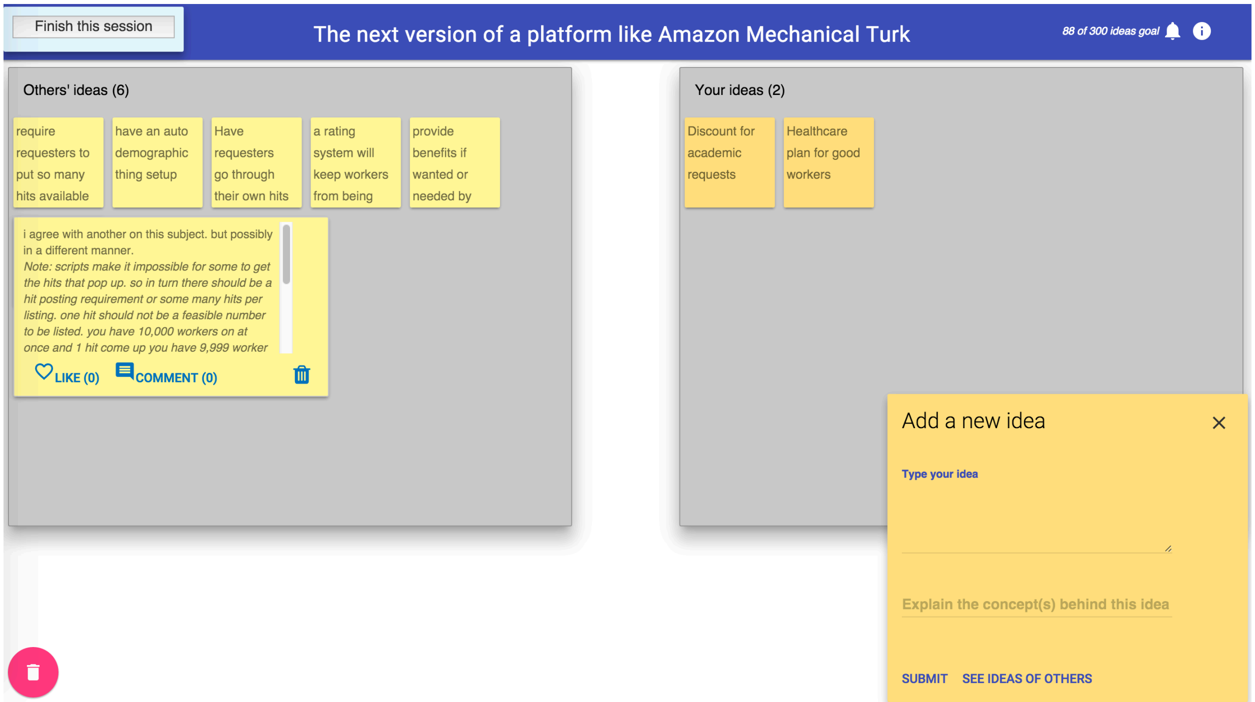


Figure 4.3: The interface for the *Single-task* condition of Study 2. The requested ideas of others are automatically placed on the left pane while participant's own ideas appear on the right pane. Participants could not move ideas around by dragging.

- *Single-task*: Participants used a more conventional system without an integrated whiteboard (Figure 4.3). They could request to see ideas of others by clicking on a "SEE IDEAS OF OTHERS" button (bottom right of Figure 4.3). The system then presented a set of three ideas sampled randomly. As with the other design, if there were no more unseen ideas, the system asked the user to request ideas again later.

Task

Participants, who were recruited via Amazon Mechanical Turk (MTurk), generated ideas in asynchronous groups of 6–12. Each group was prompted to generate ideas for one of two prompts: 1) features for the next version of a micro task market platform like MTurk (*New features*), and 2) new tasks that can be posted on a microtask market (*New tasks*). We designed these tasks such that our participants would have the relevant domain expertise and the motivation to generate novel and valuable ideas. We gave participants freedom to choose when to start their idea generation session so not all participants had to generate ideas at the same time. Participants could complete one, two or three idea generation sessions with a mandatory break of at least 15 minutes between sessions. We set up the task this way to simulate real collaborative asynchronous idea generation platforms where contributors may revisit the platform to contribute more ideas at different time. Early arrivers might have had a different experience from those who started later because they saw different compositions of ideas.

Procedure

Before starting the first idea generation session, participants answered a demographics survey. Then participants went over the tutorial of the system and completed a practice task. Following insights from prior UI evaluations on MTurk [**?**], the practice task required participants to use each major feature of the system at least once before they could proceed to the main task. For each idea generation session, participants spent at least 12 minutes on generating ideas. At the end of the session, they answered survey questions about their experience in that session. They were required to wait at least 15 minutes before starting another session. If they chose to do the next session, the system would bring them back to the saved workspace where they ended the prior session.

Participants

We recruited 80 participants via MTurk to generate ideas. We limited recruitment to workers who had completed at least 1,000 HITs with approval rate greater than 95%. After seeing some participants' comments on grammatical errors of submitted ideas in the first two groups, we limited recruitment to U.S. residents (54 participants) for the rest of the experiment. Participants were paid \$2.00, \$3.50 or \$5.00 depending on whether they completed one, two or three ideation sessions.

Out of 80 recruited participants, 55 participants finished at least one ideation session; 23 participants dropped out of the experiment during the tutorial session, and 2 participants started but did not finish the first session. We only included the participants who finished at least one session in our analysis. The participants were randomly assigned to six different groups as summarized in Table 4.1.

27 participants (87%) in the *Single-task* condition finished all three sessions, 2 participants finished only two sessions and 2 participants finished only one session. 16 participants (67%) in the *Integrated* condition finished all three sessions, 2 participants finished only two sessions and 6 participants finished only one session. On average participants in the *Single-task* condition completed 1.81 sessions compared to 1.42 sessions in the *Integrated* condition. This difference was marginally significant ($\chi^2(1, N = 55) = 3.632$, $p = 0.0567$)

Measures and analysis

We compared the creative output of participants in the two conditions on the following measures.

- *Number of submitted ideas per participant*
- *Diversity of submitted ideas*: We used the same diversity measure as in [Siangliulue et al., 2015a]. Specifically, for each group, we randomly sampled 50 pairs of submitted ideas (300 pairs for 6 groups). We recruited 58 independent MTurk workers to rate similarity of pairs of ideas on a scale of 1 (not at all similar) to 7 (very similar). Each rater rated 25 pairs of ideas from the experiment. To ensure that the workers understood the task, they also rated 4 practice pairs that were rated as very similar or very different by one of the authors. Each pair of ideas was rated by 3–4 raters. We normalized (i.e.,

converted to z-scores) the ratings—including those of practice pairs—within each rater prior to aggregating the results. We flipped the sign of z-scored similarity ratings to derive diversity scores of a pair.

- *Creativity of submitted ideas*: We used the same creativity measure as in [?]. For each group, we randomly sampled 50 submitted ideas (300 ideas for 6 groups). We recruited 58 independent MTurk workers to rate ideas on two scales: novelty (1= not at all novel, 7 = very novel) and value (1 = not at all valuable, 7 = very valuable). Each rater rated 25 ideas. Each idea was rated by 4–5 raters. As before, we converted worker ratings into z-scores prior to analysis.

To compare user experience between the two systems, we collected participants' subjective responses (reported on a 7-point Likert scale) to questions that related to the following three aspects of their ideation experience:

- *Perception of helpfulness of ideas of others as selected by the system* (4 questions)
- *Perception of helpfulness of the system* (3 questions)
- *Mental effort and task difficulty* (2 questions)

We list the actual survey questions in Table 4.2. Noting that most participants finished either just one or all three sessions, we report the survey results after the first and the third sessions.

We also asked the participants in the *Integrated* condition to answer a separate set of 7-point Likert-scale questions related to their experience of organizing ideas on the whiteboard; Q10: *"Organizing ideas on the whiteboard helped me generate ideas. (1 Strongly disagree - 7 Strongly agree)"* and Q11: *"Organizing ideas on the whiteboard got in the way of generating ideas. (1 Strongly disagree - 7 Strongly agree)"*

We used analysis of variance for analyses involving Number of submitted ideas, Diversity of submitted ideas and Creativity of submitted ideas. We used ordinal regression for all

Task	Group	Condition	Number of participants	Number of sessions	# of generated ideas
Features for AMT	G1	Integrated	8	22	58
	G2	Single-task	12	35	217
	G3	Integrated	6	16	91
	G4	Single-task	9	26	95
New types of HIT	G5	Integrated	10	20	160
	G6	Single-task	10	26	143

Table 4.1: Number of participants, sessions and submitted ideas in each group.

analyses involving Likert-scale responses. We also used ordinal regression to compare the number of sessions completed under the two conditions.

A lack of statistically significant result does not constitute valid evidence for the lack of actual difference. Because we wish to demonstrate a lack of *substantial* differences in the quality of the experience between *Integrated* and *Single-task* conditions, we also computed effect sizes (Cohen's d) for all subjective measures and some of the performance measures. As is customary, we interpret effect sizes between 0.2 and 0.49 as small, between 0.5 and 0.79 as moderate, and those larger than 0.8 as large [Cohen, 1977]. If our goal were to demonstrate the presence of statistically significant differences, we would have adjusted the p-values to account for the fact that we conducted multiple statistical comparisons based on data from a single experiment [Shaffer, 1995]. Given that our goal is the opposite, we report raw p-values throughout.

Results

No substantial difference in the number and diversity of examples seen. On average, the *Single-task* participants requested 33.8 ideas (SD=26.77), while the *Integrated* participants requested 21.6 ideas (SD=19.97). This difference is not significant ($F(1, 53) =$

3.4691, $p = 0.0681$). The average diversity scores of seen example sets were 0.074 (SD=0.46) in the *Single-task* conditions and -0.055 (SD=0.55) in the *Integrated* condition. This difference was small ($d = 0.26$) and not statistically significant $F(1, 58) = 0.9586$, $p = 0.3316$).

To derive the diversity score of examples we used a method analogous to the one used to compute the diversity of submitted ideas: We first randomly sampled 10 sets of seen examples from each group (60 sets for 6 groups). We recruited 30 independent MTurk workers to rate similarity of the 177 pairs of ideas on a scale of 1 (not at all similar) to 7 (very similar). Each rater rated up to 30 pairs of ideas. Each pair of ideas was rated by 5 raters. We normalized (i.e., converted to z-scores) the ratings within each rater prior to aggregating the results. We flipped the sign of z-scored similarity ratings to derive diversity scores of a pair. For each example set, we calculated the diversity score of an example set as the averaged pairwise diversity scores of ideas in that set.

No substantial difference in productivity. The average number of ideas submitted per session by a participant in the *Single-task* condition was 5.34, while the average number in the *Integrated* condition was 5.30. This difference was neither substantial ($d = 0.013$) nor significant ($F(1, 53) = 0.0022$, $p = 0.9626$).

No substantial difference in the diversity of submitted ideas. The average diversity score of submitted ideas in the *Single-task* condition was 0.132, while the average diversity score of submitted ideas in the *Integrated* condition was 0.105. This difference was neither substantial ($d = 0.042$) nor significant ($F(1, 299) = 0.1311$, $p = 0.7176$).

No substantial difference in the creativity of submitted ideas. The average novelty score of submitted ideas in the *Single-task* condition was 0.030, while the average novelty score of submitted ideas in the *Integrated* condition was -0.030. This difference was neither substantial ($d = 0.099$) nor significant ($F(1, 299) = 0.7309$, $p = 0.3933$).

The average value score of submitted ideas in the *Single-task* condition was 0.028, while the average value score of submitted ideas in the *Integrated* condition was -0.028. This difference

Measure	Question	Session No.	ST (Mean)	INT (Mean)	p-value	Effiect size
Perception of helpfulness of ideas of others	Q1: On average, the ideas of others that you saw were boring(1) - interesting(7)	1	5.58	5.63	0.9221	-0.0299
		3	5.67	5.19	0.4708	0.3110
	Q2: Seeing ideas of others helped me come up with better ideas. Strongly disagree(1) - Strongly agree(7)	1	5.00	5.29	0.6744	-0.1533
		3	5.26	4.56	0.1954	0.3888
	Q3: Seeing ideas of others helped me come up with more ideas. Strongly disagree(1) - Strongly agree(7)	1	4.90	5.21	0.6819	-0.1638
		3	5.07	4.69	0.4715	0.2277
	Q4: Seeing ideas of others helped me get unstuck. Strongly disagree(1) - Strongly agree(7)	1	4.90	5.21	0.6861	-0.1544
		3	5.11	4.50	0.2233	0.3224
Perception of helpfulness of the system	Q5: The system gave me a sense of what ideas other people were exploring. Strongly disagree(1) - Strongly agree(7)	1	5.68	5.33	0.4304	0.1842
		3	6.04	5.50	0.1079	0.3842
	Q6: The system helped me keep track of how my ideas related to those of others. Strongly disagree(1) - Strongly agree(7)	1	5.39	5.29	0.6075	0.0513
		3	5.89	5.44	0.2971	0.2769
	Q7: Seeing ideas of others gave me a good sense of the range of possible solutions to this challenge. Strongly disagree(1) - Strongly agree(7)	1	5.55	5.17	0.3873	0.2191
		3	5.89	5.63	0.2427	0.2042
Mental effort and task difficulty	Q8: How much mental effort (e.g., searching, remembering, thinking, deciding) did the task take? Low mental effort (1) - High mental effort (7)	1	5.52	6.21	**0.0452***	**0.6058***
		3	6.07	6.19	0.5962	-0.0851
	Q9: How easy or difficult was this task? Very easy (1) - Very difficult (7)	1	3.81	4.42	0.1469	-0.3763
		3	5.37	5.69	0.4738	-0.1952

Table 4.2: Summary of subjective responses after session 1 and after session 3. ST stands for *Single-task* and INT stands for *Integrated*. Participants rated the INT condition as demanding significantly more mental effort than the ST condition. We used Cohen's d to capture effect size.

was neither substantial ($d = 0.096$) nor significant ($F(1, 299) = 0.6843$, $p = 0.4088$).

Participants in both conditions perceived the system-selected ideas of others as similarly helpful. Questions Q1 to Q4 in Table 4.2 measured the participants' perception of the usefulness of the ideas of others selected by the system. We found no significant differences in perception of helpfulness of ideas of others between the *Single-task* and the *Integrated* condition and none of the effect sizes was larger than small. We reported the p-values and effects sizes in Table 4.2.

Participants in both conditions perceived the system as similarly helpful. Question Q5 to Q7 in Table 4.2 measured the participants' perception of the usefulness of the ideation system. We found no significant difference in perception of helpfulness of system between the *Single-task* and the *Integrated* condition and none of the effect sizes was larger than small.

The whiteboard interface initially demands more mental effort. Question Q8 and Q9 in Table 4.2 measured the participants' perception of mental effort required to do the task and the difficulty of the task. We found no significant difference of task difficulty between the *Single-task* and the *Integrated* condition and none of the effect sizes was larger than small.

However, after completing the first session, participants in the *Integrated* condition reported significantly higher mental effort than in the *Singled-task* condition ($p = 0.0452$) and this difference was moderate in magnitude ($d = 0.6058$). However, this difference was no longer present after session 3, suggesting that the system became easier to use once participants gained some practice with it.

Organizing ideas on the whiteboard helps in generating ideas and does not get in the way. The level of organization varied across participants in the *Integrated* condition (Figure 4.4). On average, a participant put 21.2 ideas on the board (SD=15.35) and formed 4.79 clusters (SD=3.52).

The responses to the 7-point Likert scale questions for participants in the *Integrated* condition show participants found that organizing ideas helped them generate ideas (Q10, session 1: M=5.17, SD = 1.69, session 3: M=4.69, SD=2.21) and that it did not get in the way of generating ideas (Q11, session 1: M=2.50, SD = 1.82, session 3: M=3.13, SD=2.31).

When further prompted to explain how organizing ideas helped them generate ideas, participants stated that organizing ideas helped them "avoid repetition, and build off of previous ideas" [P25] and "[give] a clear picture of how things were grouped and [help] brainstorm more [ideas] based on grouping" [P51]. When further prompted how organizing ideas got in the way of generating ideas, most participants either did not provide a response or stated that the activity didn't get in the way of idea generation. One participant commented that she "did spend a bit of time organizing things instead of generating ideas. But it still helped in other ways" [P31].

The sparsity of similarity matrix (from direct human judgement) increases with the number of generated ideas. Participants in the *Integrated* condition organized both ideas of others (10.6 ideas on average) and their own (11.5 ideas on average). The proportion of idea pairs that actually received human judgements in the *Integrated* group varied from 0.28 (*G5*, 160 ideas) to 0.53 (*G1*, 58 ideas). The median of number of human judgements for each pair was 1 for all groups in the *Integrated* condition. The sparsity naturally increases with the number of ideas as the size of the similarity matrix grows quadratically with the number of ideas. As we will see in the next section, this sparsity does not significantly impact the quality of the resulting semantic model.

4.5.2 Study 3: Evaluating Model Quality Using Data from the Integrated Crowdsourcing Approach

Study 2 demonstrated that the whiteboard organization successfully integrated the secondary task of semantic judgment into the primary task of idea generation in a seamless fashion. But are these semantic judgments sufficient for building an accurate semantic model? In this study, we evaluated the accuracy of the integrated semantic modeling approach by comparing an *Integrated idea map* (Figure 4.5) generated by the system for one of the *Integrated* groups from Study 2, to one generated using a previously-validated method [Siangliulue et al., 2015a] that relies on outsourced crowd workers. We will refer to this comparison semantic model as the *Outsourced idea map*.

To generate the *Outsourced idea map*, we followed the procedure described in [Siangliulue et al., 2015a]. Specifically, for the 91 ideas generated by participants in the selected group, we posted 40 MTurk tasks for workers who had not done the idea generation task to collect 1,000 responses about similarity relationship between ideas. Each worker completed a series of triplet similarity comparison tasks: “is idea A more similar to idea B or C?” [Tamuz et al., 2011]. We used an active learning heuristic to sample the questions to ask to maximize expected information gain per question [Tamuz et al., 2011]. We then used t-Distributed Stochastic Triplet Embedding (t-STE) [van der Maaten and Weinberger, 2012] to generate an *Outsourced idea map* from these responses. Although the *Integrated* and the *Outsourced* idea maps were generated from different forms of human input (spatial arrangements in the *Integrated* condition and triplet comparisons in the *Outsourced* condition), the algorithms used to aggregate the results (t-SNE [Van Der Maaten, 2014] in the *Integrated* condition and t-STE [van der Maaten and Weinberger, 2012] in the *Outsourced condition*) are both based on the same mathematical insights and should yield results with closely comparable characteristics. Thus, the key question at hand is whether collecting implicit semantic judgments from an integrated secondary task yields data of sufficient coverage and quality to build a semantic model that is at least as accurate as building a model from explicit semantic judgments collected from the external workers.

Measures and analysis

To compare the two idea maps, we measured each map against a standard baseline for comparison, which is a set of pairwise similarity ratings between ideas generated by independent human judges [Dow et al., 2010]. This similarity rating method yields accurate assessments of pairwise similarity among ideas and serves as an excellent gold standard. It is not a scalable mechanism for constructing semantic models in the first place, however, because the number of pairwise comparisons it requires grows quadratically with the number of ideas.

To obtain these independent similarity ratings, we posted 66 MTurk tasks to recruit workers (who have not previously participated in any of our other studies) to rate similarity of 550 pairs of ideas, randomly sampled across all participants, on a scale from 1 (not at all similar) to 7 (very similar). We provided a rubric with example pairs of ideas and their desired ratings. Each rater assessed 29 pairs of ideas, four of which were examples of pairs of ideas that we showed in the rubric (so that we could see if they paid attention to the instructions). Each pair of ideas was rated by at least three raters. We standardized (i.e., converted to z-scores) the ratings for each rater prior to aggregating the results. After excluding 2 workers whose answers to the rubric questions indicated that they were not paying close attention to the task, we were left with 1,725 similarity ratings.

We then computed the correlations between the human similarity ratings and the pairwise distances among ideas from each idea map. To test for potential statistical difference between the two correlations, we transformed the correlations into z-scores using Fisher's r-to-z transformation.

Results

We found a significant correlation (Spearman correlation, $\rho = -0.4848$, $p < .0001$) between distances from the *Integrated idea map* and the human similarity ratings, and a significant correlation (Spearman correlation, $\rho = -0.3878$, $p < .0001$) between distances from the *Outsourced idea map* and the human similarity ratings. Note that map distances capture *differences* among ideas while the participants were asked to assess similarity, so the negative correlation coefficient is the desirable outcome.

After transforming the correlations using Fisher's r-to-z transformation, we found the correlation between the *Integrated idea map* and human ratings to be significantly larger in magnitude than the correlation between the *Outsourced idea map* and human ratings ($z = 1.99$, $p = 0.046$). In other words, our proposed approach resulted in an idea map that better modeled the actual semantic relationships among the idea than the previous method [Siangliulue et al., 2015a] that relied on mass outsourced human computation tasks.

4.6 IdeaHound: Creativity Interventions Enabled by Real-time Semantic Modeling of Generated Ideas

Equipped with the capability to derive a computational model of semantic distances among contributed ideas, we have built IdeaHound, a system for collaborative ideation at scale. IdeaHound serves as a step towards our end goal of improved large-scale collaborative ideation. IdeaHound includes three creativity interventions enabled by the availability of a semantic model of generated ideas. These interventions are illustrated in Figure 4.6 and described here:

4.6.1 Diverse Inspirational Examples

When a user requests to see ideas of others, the system consults the global idea map and selects a set of three ideas that the user has not seen before. The requested ideas appear in the Others' ideas pane in the workspace (Figure 4.1B). Two of these ideas are substantially different according to the idea map (i.e., the distance between the two ideas on the map has to be greater than a specified threshold). The third idea is selected randomly from a pool of ideas that has been placed on none or the whiteboards. This procedure balances the need to collect judgements on newly contributed ideas and the need to present the users with ideas that are known to be substantially different from each other.

4.6.2 Similar Ideas Lookup

A user can request ideas similar to a particular idea by clicking on a request for similar idea button for that idea (Figure 4.1E). The system then consults the map to locate up to three ideas that are close to that idea (i.e., the distances between the ideas and the query idea do not exceed a specified threshold). The set of selected similar ideas will appear next to the query idea on the whiteboard (as in Figure 4.6b).

4.6.3 Visualization of the Solution Space

IdeaHound provides users with a visualization of an idea map (Figure 4.1F). The visualization shows dots on the map, each dot representing an idea. Ideas that are rendered close to each other are judged to be similar to each other. The system clusters ideas and shows a short text for the ideas that are centers of clusters to give user a quick overview of the space without cluttering the display with too many labels. Users can infer how much each part of the solution space has been explored by looking at the number of ideas in that area. They can zoom in to get a closer look at a particular region or zoom out to see an overview. The ideas submitted by the user are rendered in a different color from ideas by others to help the user see their contributions in context and decide on which direction to pursue next.

4.7 Study 4: Initial Evaluation of IdeaHound

To gauge the effectiveness of introduced interventions (and by extension, the usefulness of the semantic model produced by our integrated crowdsourcing approach), we ran a preliminary qualitative study to investigate how people use IdeaHound. This study is complementary to Study 3. While Study 3 verified the accuracy of the semantic model, Study 4 aims to provide a proof-of-concept demonstration that the semantic model generated with our approach can in fact support beneficial creativity interventions. The focus of this study is on users' experience and perception of the creativity interventions supported by the semantic model. Because we designed the study to simulate the early stages of an ideation process, we disabled looking up of similar ideas—an intervention that we hypothesized to be particularly useful in later stages of the ideation process. Thus, the focus of Study 4 is on the *Diverse inspirational examples* and *Map visualization of ideas in a solution space* interventions.

4.7.1 Participants

We recruited 7 participants (4 female) aged 18 to 32 through a call for participation sent to Harvard University students' mailing-lists. Participants were compensated $15 for taking part in the study.

4.7.2 Task

Participants generated ideas for April Fools pranks for their university. All participants worked as part of the same team. That is, they could see each others' ideas on IdeaHound.

4.7.3 Procedure

Each participant was given a link to access their workspace for the prank ideation task on IdeaHound. They then used IdeaHound to generate prank ideas in two 10-minute sessions at their own pace over the course of two days before the scheduled time for their individual in-person interviews. During the interview session, participants generated a few more ideas while thinking aloud for 5–10 minutes. They then filled out a short survey on their experience, and talked with the researcher about their experience and creative process. The entire interview session lasted about 30 minutes.

4.7.4 Results

Participants generated 115 ideas. On average, participants found the system somewhat helpful in helping them find inspirations from ideas of others (M=4.71, SD=1.74; 1 = not helpful and 7 = very helpful) and come up with ideas (M=4.57, SD=1.99; 1 = not helpful and 7 = very helpful).

Organizing ideas on the whiteboard

Six of the seven participants used the whiteboard to organize ideas. In the survey, five participants reported that they found the virtual whiteboard to be the most useful aspect of the system. Although the degrees of idea organization varied, participants who did not organize ideas as much reported that they would have organized ideas more if they had been more invested in the task and had had more time.

Participants reported organizing ideas on the whiteboard as a way to "construct [their] mind map" [P2] and establish landmarks to come back to later [P4]. Organizing ideas on whiteboard helped them see relationships between ideas (n=4), detect patterns of emerging ideas (n=4), and to kill time while thinking about new ideas (n=2). One participant [P1] reported that they did not use the whiteboard to organize ideas because looking at others' ideas or his old ideas distracted him.

Getting inspired by seeing diverse ideas sampled from the computational model

Most participants found seeing ideas of others helpful in their idea generation process. They reported building on the ideas of others and they liked to "look at others' ideas for inspiration" [P3]. P2 commented that seeing ideas of others was especially useful when he ran out of ideas. None of the participants found provided ideas of others repetitive.

However, not everyone found seeing other people's ideas helpful. One participant [P1] did not use the example request features because he likes to start generating ideas from a "blank slate" without external influences.

Reading the idea map visualization

Participants had mixed reactions to the idea map visualization. Participants used the idea map visualization to get a quick overview of ideas submitted by others (n=5), to explore many different alternatives proposed by others (n=2), and to kill time while thinking about new ideas (n=4). They also used the idea map visualization to detect patterns of the solution space [P1] and discover underexplored part of the solution space for both individual [P2, P4, P5]. P2 commented that he tried to “look for space where his ideas are not located”. Similarly, P4 and P5 stated that they tend to look at the area of the map with few ideas.

However, participants also pointed out limitations of the current version of the idea map visualization. Participants sometimes had a hard time seeing the connections between ideas that were placed close to each other on the map and expressed interest in getting an explanation of the relationships between ideas. They also mentioned that it was tricky to select the ideas that were not centers of the clusters because the size of the dots and they wished the idea map visualization would allow them to open the detail windows for more than one idea at a time.

4.8 Discussion

4.8.1 Integrating Idea Generation and Organization Into a Single Activity

Results of Study 1 demonstrated that members of volunteer communities may not always be motivated to perform work that is necessary for the good of the community, but which is perceived as tedious and as detracting from the primary interest of the community. In our case, people who were intrinsically interested in contributing novel ideas were not motivated to evaluate ideas generated by others. We have thus created an alternative interface for idea generation, one that seamlessly integrated evaluation of ideas with the primary task of idea generation.

The results from Studies 2 and 3 show that our integrated approach can model semantic relationships between ideas more accurately than a previously validated crowdsourced approach [Siangliulue et al., 2015a] with minimal impact on users' ideation experience. Although participants initially reported exerting higher mental effort with the *Integrated* system than with the conventional *Single-task* system, as the participants acclimated to the novel interface over subsequent sessions, the difference in mental effort nearly disappeared. Additionally, participants in the *Integrated* condition did not think that organizing ideas detracted from their primary task of generating ideas. This is in contrast to the results from Study 1, which demonstrated that people were not willing to evaluate ideas of others if they perceived it as an additional task. Consistent with our initial formative studies, the results from Study 4 also suggested that organizing ideas on the whiteboard *helped* idea generation by encouraging the users to make sense of the solution space upfront. A longer study could help verify whether this is the case.

Meaningfulness of clusters We expected the clusters that users generate to be meaningful because users organize ideas in IdeaHound only when it is helpful to them. However, during our formative studies, we observed that not all clusters were of equal quality. In some clusters, it was unclear why the ideas were grouped together. Introducing an affordance for adding explicit labels to clusters helped reduce this problem. Yet, a small fraction of clusters in Study 2 were not labeled. When a cluster was unlabeled, it was not always immediately clear how to derive meaning from it. Excluding all unlabeled clusters from the input to our model might improve the quality of the models by reducing noise, but it might also decrease the accuracy of the models by taking away data. Our initial experiments, in which we manually flagged and removed "noisy" clusters, did not substantially impact the quality of the resulting models. However, we plan to more systematically investigate mechanisms that can help identify and filter non-meaningful clusters to further improve the quality of the resulting idea map.

Scalability While we only showed the viability our approach in small groups of 6–10 people, this approach should also be applicable for larger ideation groups. The amount of human input required by our approach to create a semantic model of the solution space grows linearly in the number of ideas (as explained in [Siangliulue et al., 2015a]). In our method, because ideators are also organizers, the amount of input provided to organize ideas grows at the same rate as the number of ideas, so we expect no computational barriers for our system to scale. Two of our interventions (diverse inspirational examples and ability to lookup similar ideas) will not be affected negatively by the size of the community, but the idea map visualization will need to be revised so that it is still readable even if thousands of ideas are present.

4.8.2 Creativity Interventions

One might wonder why the *Integrated* intervention in Study 2 did not improve creative performance, as the results of prior work [Siangliulue et al., 2015a] would predict. A closer inspection of the example sets presented to the participants in the *Integrated* condition reveals that they were not significantly more diverse than those in the *Single-task* condition. The difference may be attributable to the way we sampled the pairs of "diverse" examples from an idea map: our algorithm first picked an idea at random from the idea map and then searched for another idea that the model predicted to be maximally dissimilar to the first. But because mundane ideas are, by definition, substantially more prevalent than unusual ones, the first randomly selected idea was almost always fairly mundane. Given that the third idea was chosen at random from among the most recently-generated ideas, this sampling approach resulted in sets of inspirational ideas that were not substantially different from those picked entirely at random. A better approach, will be to first randomly select distinct regions on the idea map (independently of the density of ideas in those regions, thus not privileging common ideas) and then sample an idea from each of these regions.

The most novel intervention we tested was the idea map visualization, which presented a succinct synthesis of the ideas explored by the community so far. Participants in Study 4 found this idea map visualization useful in providing an overview of the solution space. Some participants [P4, P5] used the visualization to identify underexplored parts of the solution space and to decide how best to contribute to the group effort. Thus, an idea map visualization can act as a guide to coordinate group ideation effort by directing people to explore different parts of the map to avoid redundant work.

The results of Study 4 also suggest ways to improve the interface of the idea map visualization. Specifically, participants sometimes did not understand why certain ideas appeared close to each other on the idea map visualization and would have liked to see explanations of the semantic relationships implied by the visualization. This finding is likely explained by the fact that different people appeared to construct different mental models of the emerging solution space. In Studies 2 and 4 we repeatedly observed that different participants grouped the same ideas differently. This observation is consistent with prior findings [André et al., 2014]. While this still allowed the algorithm to create a computational model that captured meaningful semantic relationships among ideas, it suggests that a grouping that is intuitive to one participant may be surprising to another. In the future, we will leverage the labels participants attach to the clusters they create. As these labels explicitly reveal shared semantics among a group of ideas, they may be the right vocabulary with which to communicate the rationale behind different clusters on the idea map visualization.

4.9 Conclusion

Prior work on creative cognition and creativity support tools demonstrated that having a computational semantic model of a solution space can enable a number of interventions that demonstrably improve the number, quality and diversity of ideas people generate. In large-scale online innovation platforms, where people contribute ideas in the form of short text snippets or sketches, no prior feasible mechanism existed for creating such computational models. We contribute such a mechanism: it combines human judgements with machine learning to estimate similarity among all ideas contributed by a community. Because people were not willing to contribute subjective judgements of idea similarity when they perceived this to be a separate task unrelated to the primary activity of idea generation, we developed a novel system, called IdeaHound, which seamlessly integrates the secondary task of providing feedback on semantic relationships among ideas into the primary task of idea generation.

The results of our studies demonstrate the viability of our approach. We found that people were as willing to use IdeaHound to simultaneously generate and organize ideas as they were a conventional design that did not require organizing ideas. Furthermore, the subjective judgements implicitly collected through IdeaHound resulted in a more accurate computational model of semantic relationships among ideas than an existing approach [**?**], which relied on outsourcing the task to an external crowd.

We also show how this computational model can support creative interventions that users find useful, specifically, providing diverse inspirational examples, and providing an overview of the solution space in the form of an idea map visualization.

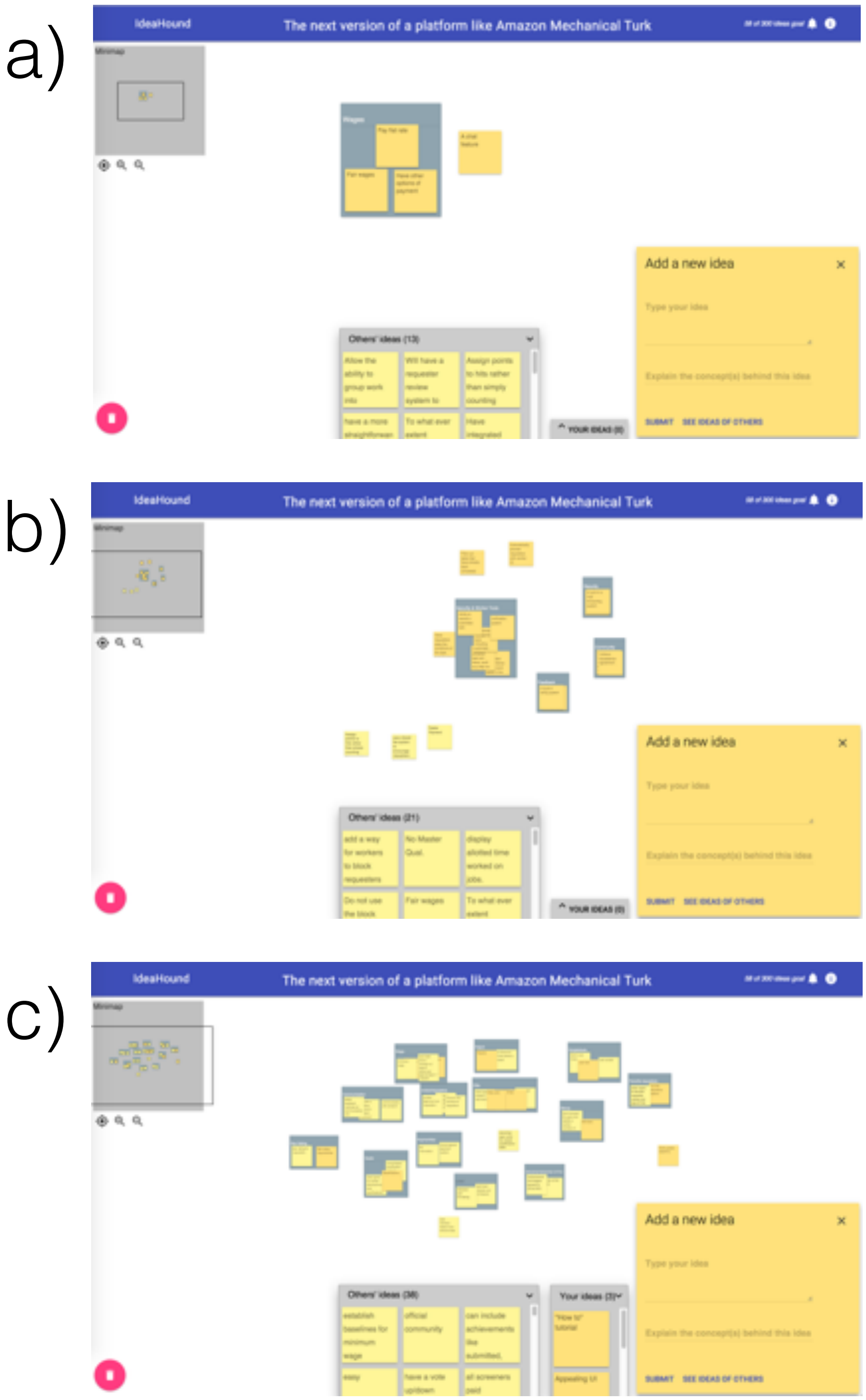


Figure 4.4: Participants engaged in organizing ideas to varying extent, ranging from making hardly any clusters (a: *P5*, *G1*), to moderate organization (b: *P6*, *G1*), to extensive organization (c: *P31*, *G3*).

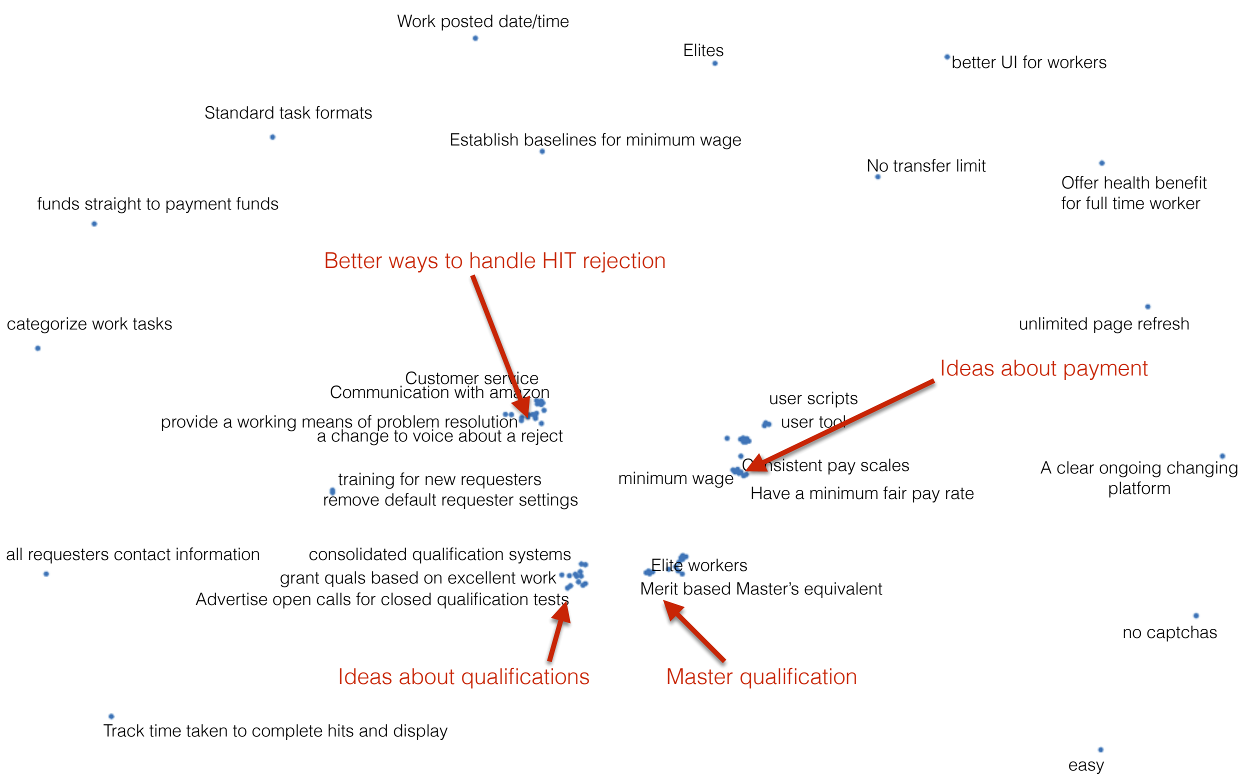


Figure 4.5: An idea map of *G3* in the *Integrated* condition, showing clusters of ideas around different topics. Isolated ideas around the edge are the ideas that either are different from other ideas or are the ideas that the system does not know much about yet.

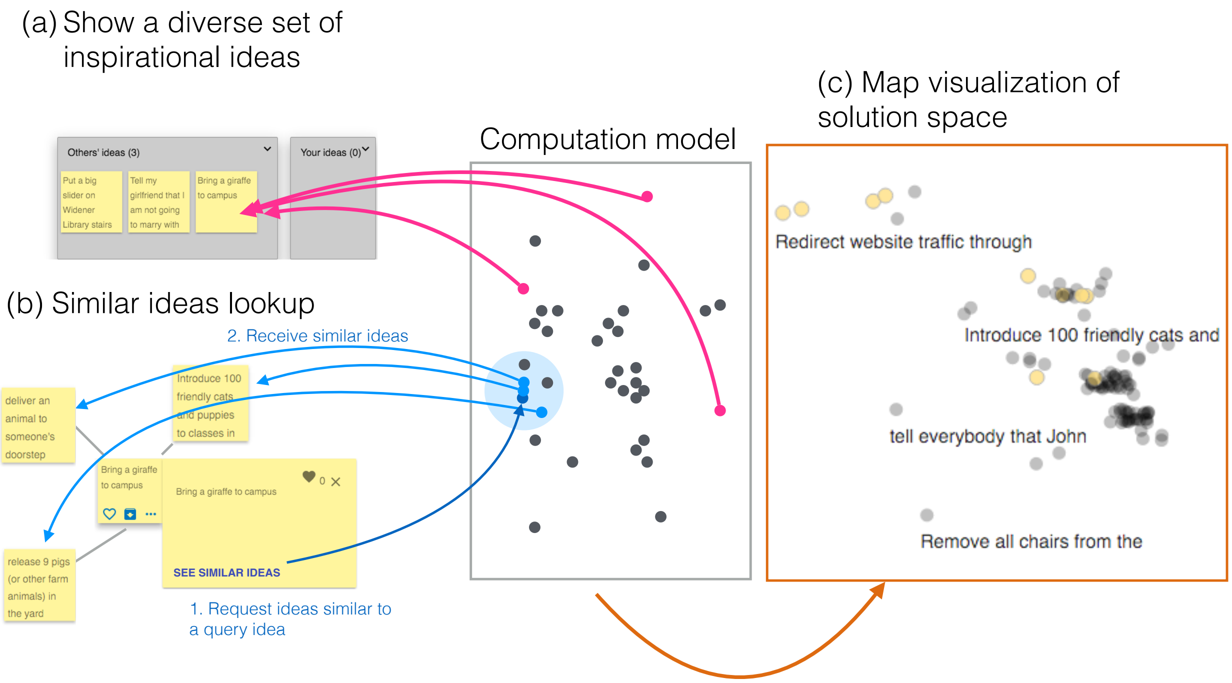


Figure 4.6: Proposed interventions to improve experience and output of idea generation task as implemented by IdeaHound . The computational model box represents show the shape of the solution space through idea instances and their relationships. (a) When a user requests to see ideas of others, the system selects a set of diverse ideas (instead of sample randomly). (b) A user can ask to see a set of ideas that are similar to a certain idea. (c) A user can get a quick overview of the solution space through an map visualization that shows their ideas and ideas of others in the solution space.

Chapter 5

Providing Timely Examples

This chapter has adapted, updated, and rewritten content from a paper at Creativity and Cognition 2015 [Siangliulue et al., 2015b]. All uses of "we", "our", and "us" in this chapter refer to coauthors of the aforementioned paper.

Emerging online creative communities with thousands of example ideas provide an important resource for creative production. But how can community members best use these examples to create new innovations? Recent work has suggested that not just the choice of examples, but also the timing of their delivery can impact creative outcomes. Building on existing cognitive theories of creative insight, we hypothesize that people are likely to benefit from examples when they run out of ideas. We explore two example delivery mechanisms that test this hypothesis: 1) a system that proactively provides examples when a user appears to have run out of ideas, and 2) a system that provides examples when a user explicitly requests them. Our online experiment (N=97) compared these two mechanisms against two baselines: providing no examples and automatically showing examples at a regular interval. Participants who requested examples themselves generated ideas that were rated the most novel by external evaluators. Participants who received ideas automatically when they appeared to be stuck produced the most ideas. Importantly, participants who received examples at a regular interval generated fewer ideas than participants who received no examples, suggesting that mere access to examples is not sufficient for creative inspiration. These results emphasize the importance of the timing of example delivery. Insights from this study can inform the design of collective ideation support systems that help people generate many high quality ideas.

5.1 Motivation and Contributions

Online creative communities—such as Quirky.com, Innocentive.com, 99designs.com—accumulate thousands of ideas contributed by their members. Because the members can see and be inspired by each other's ideas, these collections of example ideas can serve as an important resource for creative production [Chan et al., 2014]. Ideas generated by others can help innovators working on similar problems spur new concepts by broadening their notion of the design space [Herring et al., 2009, Lee et al., 2010, Ritchie et al., 2011] and allowing for reinterpretation and recombination of ideas [Herring et al., 2009, Yu and Nickerson, 2011, Marsh et al., 1996]. When viewing ideas for inspiration, innovators should pay attention to how to select examples [Lee et al., 2010, Siangliulue et al., 2015a, Kumar et al., 2013, Ritchie et al., 2011], and how to judge their quality [Herring et al., 2009]. This is especially important because exposure to other ideas is not always inspirational: people often transfer solution elements from other ideas even when those ideas are known to be of low quality [Chrysikou and Weisberg, 2005, Jansson and Smith, 1991b]. Recent research shows that even experts are susceptible to such negative effects of exposure to other ideas [Linsey et al., 2010]. Other ideas can also restrict one's understanding of the solution space, for example, by limiting one's ability to see novel uses for artifacts [German and Barrett, 2005, Maier, 1931].

Consequently, much research attention has been devoted to understanding which properties of examples are associated with inspirational outcomes. For example, research has considered how the semantic relevance [Chan et al., 2014, 2011, Dahl and Moreau, 2002], novelty [Chan et al., 2011, Agogué et al., 2013], and diversity [Doboli et al., 2014, Zeng et al., 2011, Baruah and Paulus, 2011, Siangliulue et al., 2015a] of examples influence ideation. However, one important question has received less attention: *when* should innovators look at examples?

A variety of theoretical perspectives suggest that the impact of examples on creative output not only depends on what examples are shown but also when those examples are delivered. Cognitive theories of creative ideation suggest that ill-timed examples can disrupt a person's train of thought [Nijstad et al., 2002, Nijstad and Stroebe, 2006] and that people benefit most from examples when they run out of ideas [Seifert et al., 1995, Patalano and Seifert, 1994, Moss et al., 2007]. Research on flow and interruptions also suggest that automatic example delivery can be experienced as an as interruption if not timed appropriately [Bailey et al., 2000, Bailey and Iqbal, 2008, Csikszentmihalyi, 1997], thereby harming creative performance. However, the literature lacks empirical tests of these hypotheses.

This chapter presents an empirical test on whether people benefit more from examples when they are prepared to receive them compared to seeing those same examples delivered at fixed intervals. We conducted an online ideation experiment to test two "prepared" conditions—an *On-demand* condition, in which participants determined when to see examples, and an *On-idle* condition, in which participants were automatically presented with new examples when they had been idle for a period of time. We compared these conditions against two baselines: a condition where no examples were provided (*None*) and a condition where the examples were provided at a regular interval (*On-interval*). The baseline conditions let us distinguish the effect of access to examples *per se* from the effect of timing of the delivery of examples.

Our results show that both prepared conditions outperform the baseline conditions, but in different ways. Participants who received examples on demand produced ideas that were deemed significantly more novel by evaluators compared to participants who did not receive any examples and to participants who received examples when idle. Meanwhile, participants

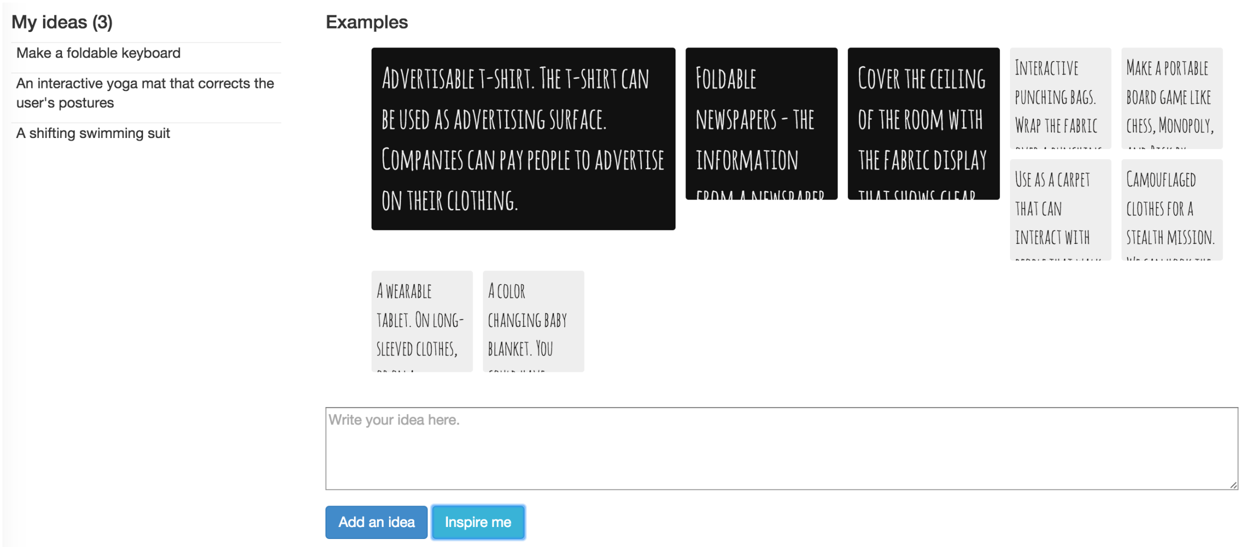


Figure 5.1: Screenshot of the ideation interface. Participants typed their ideas in the text box. After they submitted an idea, it appeared on the pane on the left. For those in the *On-demand*, *On-idle* and *On-interval* condition, examples were shown in the example grid above the idea entry box. The most recently received examples were shown in a big black box while earlier examples were in a gray box. The "Inspire me" button at the bottom was visible only to participants in the *On-demand* condition.

who received examples automatically whenever they were idle produced a larger quantity of ideas than participants in other conditions, with no significant difference in novelty compared to ideas generated by participants in either of the baseline conditions. Finally, a follow-up content analysis of the participants' ideas showed that participants who received examples on demand used examples more (i.e., borrowed/adapted more solution elements) compared to participants who received examples when idle. These results confirm that the timing of example delivery can determine the impact of examples on creative output. From a system designer's perspective, our results suggest that, instead of giving people examples in an ad hoc way, the examples should be presented at the right moment when the user is ready to make use of those examples.

5.2 Timing of Example Delivery

We explore two mechanisms for delivering examples to innovators when they are prepared to receive them. The first mechanism is to provide examples when people explicitly request them (the *On-demand* condition). This approach guarantees that the examples will be provided when people are receptive to new ideas [Friedman et al., 2003, Seifert et al., 1995, Patalano and Seifert, 1994, Moss et al., 2007]. However, people might choose suboptimal strategies for requesting examples (e.g., spending too much time looking at inspiration). People might also not be aware that they are stuck in (or biased by) old patterns of thinking [Marsh et al., 1997, Bilalić et al., 2008, Ward, 1994] and consequently fail to request examples at an opportune time.

The second mechanism automatically provides the examples when people appear to be stuck (the *On-idle* condition). We used a simple timeout mechanism: when no activity was detected in the interface for a fixed period of time, the system automatically provided a new set of examples of ideas generated by others. Prior research provides little guidance on how idle time during ideation relates to being in a "stuck" state. Therefore, we conducted a pilot study where we observed three people generating ideas in person. We looked at big time gaps between bursts of successive idea generation. Interviews with participants revealed that during these time gaps, they ran out of ideas on one thread and then started a new train of thought. We observed that these gaps tended to be approximately 30 seconds long. Thus, we decided on a fixed idle interval of 30 seconds for the *On-idle* condition. Analyses of time gaps before example requests in the *On-demand* condition of our main experiment provide further support for this choice of idle interval.

5.3 Experiment

5.3.1 Participants

We recruited 120 participants from Amazon Mechanical Turk[1] (MTurk), an online micro-labor market. Three participants did not complete the experiment and were excluded from our analysis.

We limited recruitment to workers who resided in the U.S. and who had completed at least 1,000 HITs with greater than 95% approval rate (to reduce noise from less skilled or motivated workers). Participants were paid $2.50 for their participation.

5.3.2 Task and Procedure

Each participant completed two idea generation tasks. In the first task, they had 3 minutes to generate as many alternative uses for rubber bands as possible. This was a warm-up task designed to familiarize participants with the system and with the example delivery mechanism. We did not include the data from this task in our analysis. In the second task, participants had 15 minutes to generate product ideas for an imaginary technology—a touch-sensitive "fabric display" that could render high resolution images and videos on any fabric through a penny-sized connector. We selected this task because it did not require extensive expertise to generate ideas, but yet was more similar to realistic design tasks than toy problems (e.g., alternative uses for a rubber band).

At the beginning of the experiment, each participant was randomly assigned to one of the four conditions:

- *On-demand*: Participants could request a new set of three examples whenever they wanted until they saw all available examples.

[1] http://www.mturk.com

- *On-idle*: Participants were automatically presented with a new set of three examples when they stopped typing for 30 seconds.
- *On-interval*: Participants saw a new set of three examples at the beginning of the task and on regular intervals afterward (every minute for the alternative uses task and every three minutes for the product ideas task).
- *None*: Participants saw no examples while generating ideas.

When new examples appeared, they appeared in a set of three and were shown prominently at the top of the example grid until another set of examples came. Older examples were available throughout the idea generation session, but they were less visually prominent (Figure 5.1). Before each idea generation session, all participants were informed about how and when they would have an access to a new set of examples. After finishing the second task, participants filled out a survey on their demographic information and their experience during the last idea generation session.

5.3.3 Examples

There were 9 examples available for the alternative uses task and 15 examples for the product ideas task. Examples for the alternative uses task were obtained through an Internet search. Examples for the product ideas task were obtained from a pilot round of idea generation with 12 MTurk workers generating ideas for 15 minutes each. We selected examples as follows. A trained coder (an author) evaluated the 71 potential examples for the alternative uses task and the 60 ideas collected in a pilot study of the product idea tasks. The product ideas were coded with thematic tags like "advertising" and "camouflage." We also assessed the overall quality of each idea (judging both novelty and value). We assembled sets of three ideas that comprised both high quality and diverse theme, as both example quality and diversity have been shown to improve ideation performance [Paulus and Dzindolet, 1993, Leggett Dugosh and Paulus, 2005, Nijstad et al., 2002, Siangliulue et al., 2015a].

5.3.4 Dependent Measures And Analysis

We conducted a between-subjects study with timing of example delivery (*None*, *On-demand*, *On-idle* and *On-interval*) as the sole factor.

We collected three performance measures:

- *Number of nonredundant generated ideas.* Six redundant ideas were removed by me. A sample (249 raw ideas by 29 participants) was also evaluated for redundancy by the second author of the original paper, and the reliability was high, ICC(2,2) = 0.83.
- *Novelty* of ideas as assessed by other MTurk workers (who were not participants in the ideation study). Previous work has also used MTurk workers to evaluate creativity of ideas (e.g., [Yu and Nickerson, 2011]).
- *Value* of ideas as assessed by other MTurk workers. This measure maps onto the dimensions of appropriateness (quality) and feasibility typically used in prior studies of creativity.

To evaluate Novelty and Value, each MTurk judge rated a random sample of 25–30 ideas. The evaluators were asked to read all ideas before rating them on 2 criteria, novelty and value, each on a 7-point likert scale. For novelty, we asked them to "consider how novel, original or surprising the idea is" (1–Not novel; 7–Very novel). For value, we asked them to "consider how useful the product idea is and how practical the idea sounds assuming the 'fabric display' technology is real" (1–Not valuable; 7—Very valuable).

Each of our evaluators rated a different subset of artifacts so calculating the agreement between evaluators is not feasible. However, we have evidence from a prior reliability study that this rating approach yields satisfactory reliability. Using three different types of creative artifacts, we measured how reliability improved as we increased the number of MTurk workers assessing creativity of any one idea. We found that a panel of three raters achieved inter-panel intraclass correlation coefficient (ICC) of 0.432. Most (98.6%) of our ideas in this study were evaluated by at least three evaluators.

To address potential misalignments in absolute means and variances in scores between evaluators, we first normalized each evaluator's scores into z-scores. We then averaged the normalized (z-)scores for each idea across evaluators. A 0 z-score meant that an idea was rated average, negative z-score means that the idea was rated below average on that criterion.

To illustrate, here are examples of ideas with low novelty (z-)scores:

- *"material for a hat"* (-1.88)
- *"games"* (-1.87)

While these are ideas with high novelty scores:

- *"Curtains that make it look like people are home when they are way. as part of a security system"* (1.78)
- *"Neckties - If they spill something on it at lunch, they can change the color so it blends in and don't have to worry about anyone noticing the stain."* (1.28)

Here are examples of ideas with low value scores:

- *"A wearable table. On long sleeved clothes."* (-1.83)
- *"A color changing bra that displays your favorite apps."* (-1.60)

While these are ideas with high value scores:

- *"Use as a stealth device for soldiers to get behind enemy lines."* (1.73)
- *"Provide to underfunded schools to replace their expensive projectors in classrooms."* (1.44)

Once they finished generating ideas, participants in the *On-demand* condition answered survey questions about when and why they requested examples (Table 5.1).

We also recorded timestamps when ideas got submitted and when participants saw a new set of examples. Using these timestamps, we looked at how much time passed after the latest idea submission before participants requested new examples.

5.3.5 Adjustments to the Data

There were originally 25 participants in the *None* condition, 26 participants in the *On-demand* condition, 31 participants in the *On-idle* condition and 35 participants in the *On-interval* condition. Our random assignment mechanism did not ensure balanced numbers across conditions because some MTurk workers abandoned the tasks when the conditions were already assigned, hindering accurate counting of participants in different conditions.

We filtered out the participants who either never requested examples or requested examples only once because these participants might not have understood that they could request examples or keep requesting examples more than once. This excluded 7 out of 26 participants from the *On-demand* condition. Because evaluating ideas is costly and the numbers of participants were unbalanced, we further randomly sub-sampled participants in the *On-idle* and the *On-interval* conditions so that similar numbers of participants from each condition would be used in the final analysis.

We ended up with 97 participants: 25 in the *None* condition, 19 participants in the *On-demand* condition, 28 participants in the *On-idle* condition, and 25 participants in the *On-interval* condition. These participants (along with their 1,149 ideas) constitute the final sample for our analysis.

5.4 Results

5.4.1 Providing examples at idle time led to more ideas

We observed a significant main effect of timing of example delivery on the number of ideas generated by participants (F(3,93)=3.26, $p = 0.0249$). On Average, participants in the *On-idle* condition generated the most ideas (M=13.8), followed by participants in the *On-demand* condition (M=10.94), the *None* condition (M=10.88) and the *On-interval* condition (M=8.80) (Figure 5.2). The pairwise Student's T comparisons show significant difference between participants in the *On-idle* condition and the *On-interval* condition. There was no difference between the other pairs.

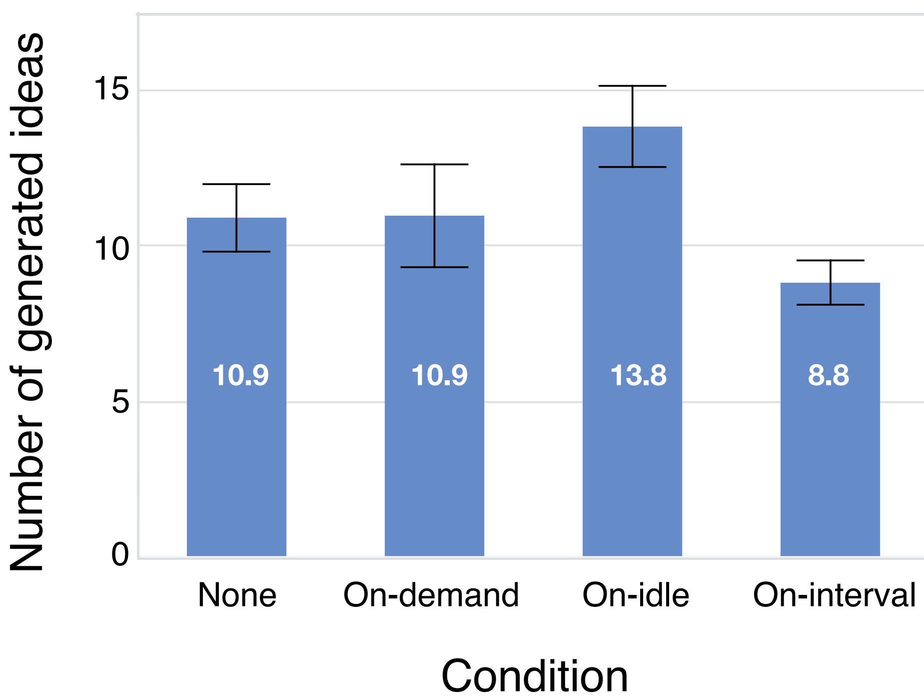


Figure 5.2: Participants in the *On-idle* condition generated significantly more ideas than participants in the *On-interval* condition. Error bars show standard error.

5.4.2 On-demand example requests led to more novel ideas

We observed a significant main effect of timing of example delivery on the average novelty of ideas (F(3,93)=4.89, $p = 0.0034$). The pairwise Student's T comparisons show that participants in the *On-demand* condition (M=0.18) generated ideas that were deemed significantly more novel than those in the *None* condition (M=-0.18) and those in the *On-idle* condition (M=-0.01). The difference between the *On-demand* condition and the *On-interval* condition (M=0.05) was not significant (Figure 5.3).

We did not observe any statistically significant differences across conditions for the average value rating of ideas (F(3,93)=1.18, $p = 0.32$).

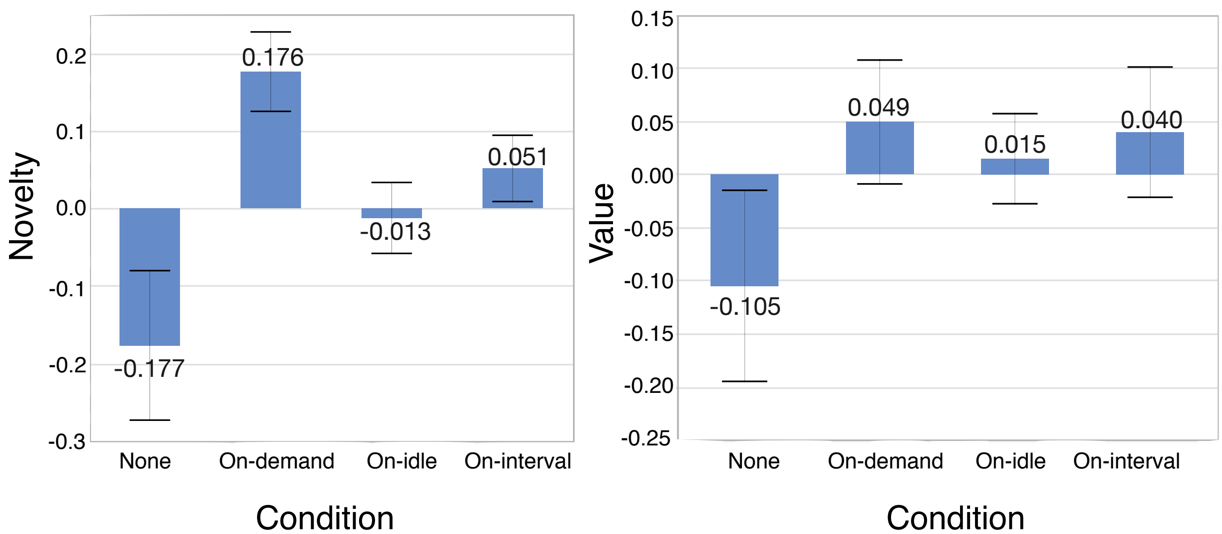


Figure 5.3: The mean novelty z-score for participants in the *On-demand* condition is significantly higher than for those in the *None* and *On-idle* condition. There is no statistically significant difference across conditions for the value scores. Error bars show standard error.

5.5 Follow-Up Analyses

We conducted two sets of follow-up analyses to address questions raised by the main findings. These analyses focused on understanding why and when participants requested examples, and exploring hypotheses about why the prepared conditions (i.e., the *On-demand* and *On-idle* conditions) had different impacts on participants' creative performance.

5.5.1 Why and when did participants request examples?

Table 5.1 summarizes the survey responses of participants in the *On-demand* condition on why and when they requested examples. The responses indicate that participants primarily requested examples when they ran out of ideas. A smaller (but still sizable) proportion of participants appeared to use an alternative strategy where they looked at examples before generating ideas.

When did you request examples?	Participants N (%)
When I ran out of ideas.	15 (78.95%)
Before I started generating my own ideas.	6 (31.58%)
In the middle of coming up with new ideas.	3 (15.79%)
When I got bored.	2 (10.53%)

Table 5.1: When did the *On-demand* participants request examples? The majority of participants said in the post-experiment that they requested examples when they ran out of ideas.

On average, participants requested a new set of examples 31.19 seconds (SD = 44.37s) after they submitted their latest ideas (excluding example requests that came before participants submitted their first idea). This average idle time suggests that our choice of 30s delay in the *On-idle* condition was reasonable.

However, inspecting these idle time distributions across the session yields a more nuanced picture (Figure 5.4). First, idle times before requesting examples tend to be shorter earlier in a session: idle times for the first and second example requests tended to be shorter than 30s. Second, there was a considerable amount of variability between participants in terms of idle times: while the mean idle time is close to 30s, participants sometimes waited more than a minute before requesting examples.

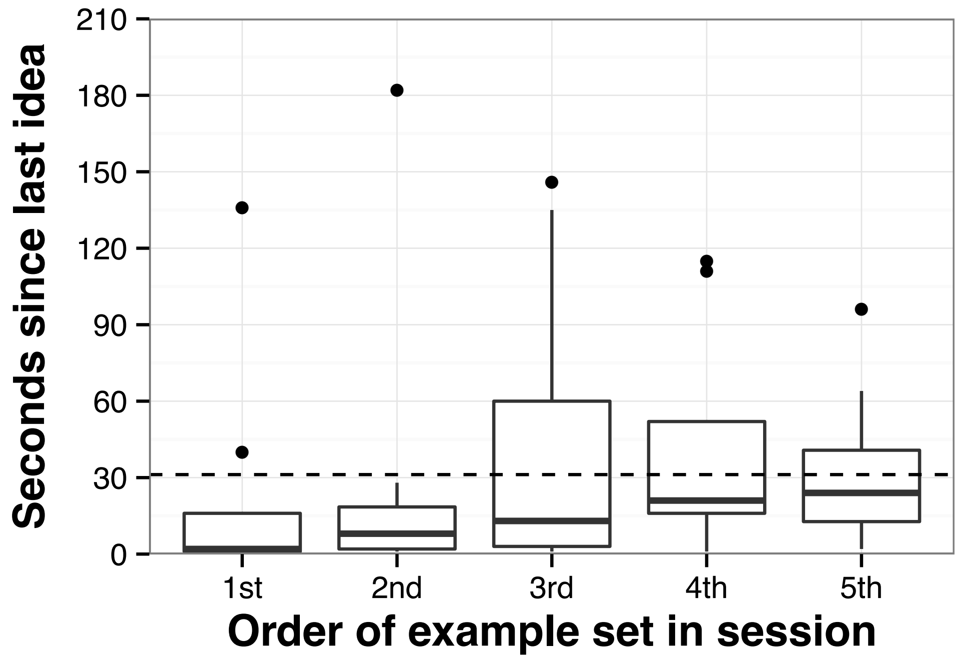


Figure 5.4: Boxplot of idle time before example request by order of example set in session. The mean time before requesting examples was 31.19 seconds. Participants were idle for shorter amounts of time before requesting first and second example sets than for third, fourth and fifth sets. Participants' idle times also varied considerably, with some participants waiting longer than a minute before requesting examples.

5.5.2 How did participants use examples?

To better understand the observed differences between the *On-demand* and *On-idle* conditions, we conducted a content analysis of the examples' impact on participants' ideas. We sampled all ideas that participants generated immediately after an example set was seen to compare against their corresponding example sets. We also included the most recent prior idea (generated within 30 seconds or less than the last seen example set) for comparison because it was common for participants to generate successive ideas within the same category or with shared functional features. In some cases, example sets were seen in succession without any ideas generated in between. In these cases, we considered the impact of the last set of examples on the next idea. This sampling procedure yielded 145 example-idea cases: 89 in the *On-idle* condition, and 56 in the *On-demand* condition. Our goal was to identify whether and how examples influenced the ideas participants generated.

The content analysis was conducted by an expert panel comprising me and the second author of the original paper. The panel separately analyzed each example-idea case to identify whether the idea appeared to be influenced by any of the examples just seen. The prior idea was included as a comparison point, since features in the idea could have plausibly been transferred/adapted from a prior idea, rather than from one or more of the examples [Nijstad and Stroebe, 2006]. We only considered features shared with examples that did not overlap with those of the prior idea. Specifically, we considered two kinds of example influence, following cognitive theories about example use [Bearman et al., 2002, Ball et al., 2004]:

1. Transfer of *structural features*, where the panel agreed that the idea appeared to contain mechanisms or functions (e.g., interactivity, simulation, tailoring displays to states of a system, sensing user states) also present in one or more of the preceding examples (and absent in the prior idea). For example, the idea *"Safety warnings from public institutions i.e. different colored flags on the highway that reflect Amber Alerts or how safe the roads are (a color co[d]ed system will be in place)."* shares the same mechanism of displaying the state of the systems or environment with *"Stuff animals with emotions. Make stuff animals out of this fabric. They can smile when hugged or make different facial expressions"*.

2. Transfer of *surface features*, where the panel agreed that the idea appeared to share application contexts (e.g., use for health/exercise, sports/games, learning/education)

and basic features (e.g., positioning on clothing/furniture) also present in one or more preceding examples (and absent in the prior idea). For example, the idea *"To have beating organs on the outside of your clothing"* shares the same domain concept—body organ—with *"Attached with a sensor to detect body heat or heart rate, the fabric can make for clothes that detect if you are stressed out or fatigued. It will display peaceful images in soothing colors when you are stressed out"*.

Structural and surface features were considered separately to examine the possibility that participants in the *On-demand* condition were generating more novel ideas by engaging in far transfer (i.e., transferring structural features but not surface features [Dahl and Moreau, 2002]). The panel also took note of the *number* of examples that appeared to have influenced the idea.

The panel was blind to condition throughout the analysis. The panel first identified a list of features considered to be structural and surface. Then, the panel analyzed each example-idea case in an iterative manner with discussions progressing until resolution was reached. Earlier coded cases were reanalyzed in light of insights gained from later cases.

Out of the 145 example-idea cases, only 4 ideas (from 2 participants from the *on-idle* condition) were identical (or nearly identical) to the given examples. For example, a participant generated idea *"A flag that changes between various nations."* when they saw an example *"A multinational flag. Instead of having more than different flags for different nations, you can save space by having one flag that rotate showing flags of various nations."* We further inspected the ideas of these two participants and found that the copied ideas made up only a small portion of their generated ideas. The panel did not count structural or surface transfers from these copied ideas.

Figure 5.5 shows transfer rates for the *On-idle* and *On-demand* conditions (averaged across participants). A simple z-test for a difference in proportions yields a significant coefficient(z=3.82, $p < .001$), indicating that transfer was observed in a statistically higher proportion of cases in the *On-demand* condition compared to the *On-idle* condition. This data suggest that participants in the *On-demand* condition used examples more often than participants in the *On-idle* condition.

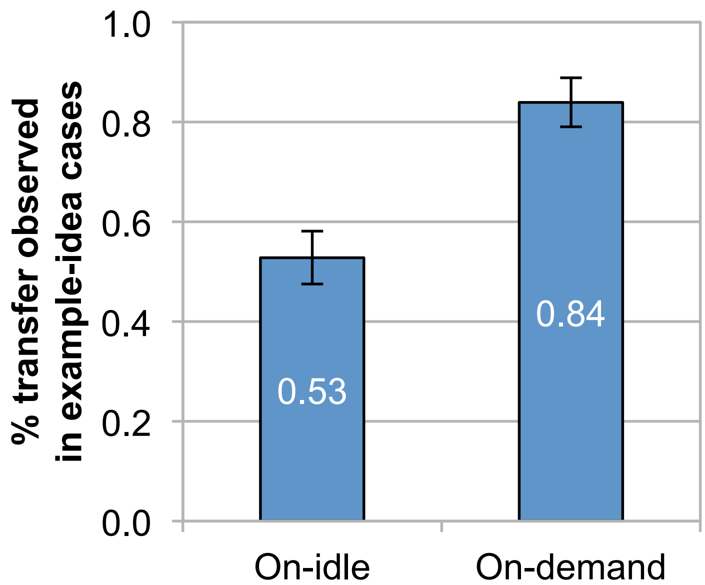


Figure 5.5: Participants in the *On-demand* condition used more examples to generate new ideas than those in the *On-idle* condition as shown in this figure where the *On-demand* participants transferred more features from examples to their ideas the the *On-idle* participants.

Analysis by type of feature transfer yielded similar results (see Figure 5.6). Transfer rates were higher for *On-demand* cases for both structural (z=2.55, $p < .05$) and surface features (z=3.68, $p < .001$). Importantly, the ratio of structural to surface transfers was similar for both conditions. These findings suggest that differences in novelty between the *On-demand* and *On-idle* conditions may be due to quantitative (i.e., more cases of examples actually influencing ideation) rather than qualitative differences (e.g., more sophisticated transfer) in how the participants used the examples.

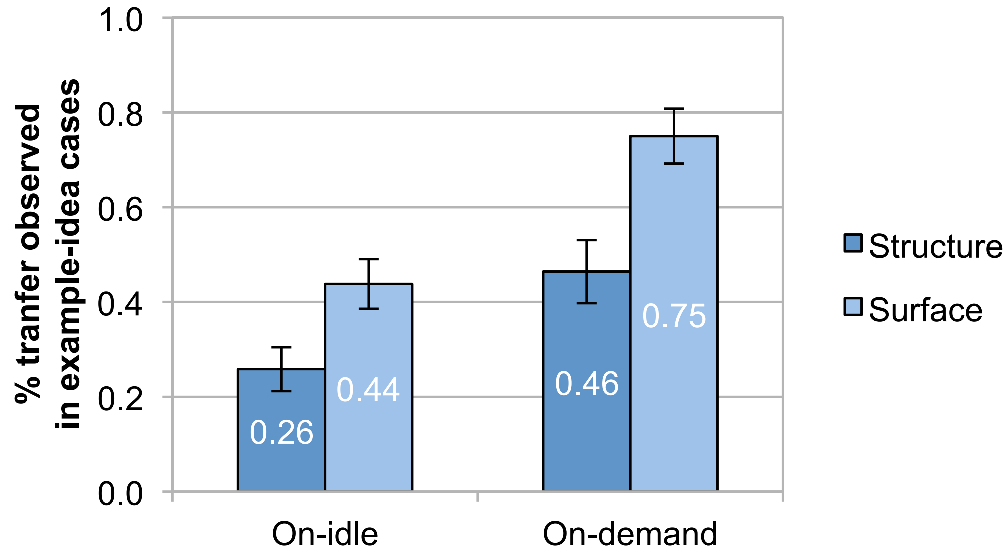


Figure 5.6: Participants in the *On-demand* condition transferred both structural and surface features from examples more often than those in the *On-idle* condition.

.

5.6 Discussion

Adding to prior work showing the importance of considering *what* examples to see, our results demonstrate the importance of carefully considering *when* to see examples. Giving participants acces to examples on demand led to the highest ratings for novelty (but did not boost productivity). Automatically presenting examples to participants whenever they were idle also benefited ideation, but only for number (and not novelty/value) of ideas. In contrast, participants who received examples at regular intervals produced the fewest ideas (even fewer than participants who saw no examples at all). We now unpack these findings in more detail and draw out their implications for cognitive theories of creativity and the design of platforms for collaborative inspiration.

5.6.1 Why were on-demand and on-idle effects so different?

Why were there differences in the novelty of generated ideas between the *On-demand* and the *On-idle* conditions given that both interventions aimed to offer examples to people when they were stuck in a mental rut? One possible explanation may be related to our specific mechanism for automatically inferring when the person was stuck. Delivering examples when a person is idle for 30 seconds might be too simple or we might not have picked the right threshold time to infer the stuck moment. Our follow-up analyses of the idle timing data from the *On-demand* condition showed that the average waiting time was approximately 30 seconds, suggesting that, on average, our choice of idle interval was reasonable. Nevertheless, there was also variability in the wait times, both between participants and within sessions. While we do not believe the pattern of effects of on-idle examples is idiosyncratic to our choice of idle interval, future work exploring more nuanced idle intervals might yield more precise estimates of the size of these effects.

From a theoretical perspective, a more interesting alternative explanation might be that awareness is a key component of a prepared mind: that is, to benefit from inspirational stimuli, being stuck is not enough—you must also know that you are stuck. Theoretically, our results suggest that theories of creative insight inspiration (such as the Prepared Mind theory [Seifert et al., 1995, Patalano and Seifert, 1994, Moss et al., 2007]) should pay more attention to metacognitive factors (e.g., awareness of one's own cognitive states). Practically, interventions designed to increase metacognitive awareness (e.g., mindfulness training) may help people maximize opportunities for inspiration. Future experiments might explore if on-idle inspiration delivery combined with such interventions could match the benefits of on-demand example delivery.

Alternatively, perhaps participants in the *On-idle* condition benefited less from examples because the examples were delivered while they were still productively accessing knowledge within a given category, even if they were not typing into the system. Our example sets were diverse and would probably have required participants to switch categories in order to recombine them into new ideas. SIAM theory [Nijstad et al., 2002, Nijstad and Stroebe, 2006] predicts that switching categories requires effort, and can lead to productivity losses. Perhaps *On-idle* inspiration delivery could still be beneficial if the examples were "personalized" (e.g., coherent extensions of a user's current solution path). Such examples could activate other knowledge that is related to currently activated knowledge. Prior work has suggested that deep exploration within a category is an alternative (and often overlooked) pathway to highly creative ideas [Nijstad et al., 2010b, Rietzschel et al., 2007a]. Future work could develop novel mechanisms for real-time semantic analysis of participants' solution paths, and conduct experiments to test whether personalized inspiration could further help people benefit from inspirational examples.

Although participants in the *On-idle* condition produced ideas that were rated as slightly less novel than those generated by participants who received examples on demand, they were the most productive. This result suggests that we can prime people to produce more ideas by showing them examples when they are idle without sacrificing the novelty or value of generated ideas. This productivity gain might be explained by the fact that new examples were presented to them before they realized that they were stuck, allowing them to pursue a new train of thought sooner instead of wasting time waiting for new ideas. However, the follow-up analysis suggested that participants in the *On-idle* condition did not use examples to guide their ideation as often as the *On-demand* participants. An alternative explanation that is more consistent with the data is that the appearance of a new set of examples signaled to people that their performance was being monitored and thus nudged them to keep on working. Prior work has shown that people increase their rate of idea generation when they know their work is being watched or will later be evaluated [Weber and Hertel, 2007, Michinov and Primois, 2005, Shepherd et al., 1995]. However, there is little evidence that this increased productivity also leads to higher quality (or more novel) ideas. Indeed, people often refrain from exploring "wild ideas" if they know or perceive that they are being evaluated for their ideas, a phenomenon known as evaluation apprehension [Diehl and Stroebe, 1987, Cooper et al., 1998]. Future work that explores idea generation systems with automatic example delivery mechanisms should test this alternative explanation, and carefully consider participants' perceptions of automated support when designing such systems.

5.6.2 Did examples really help?

One important question to consider in interpreting the results is whether the examples really helped. For example, did participants in the *on-demand* condition merely copy features from the examples? Our follow-up content analysis suggests that they did indeed use examples to guide their idea generation to a greater extent than the *on-idle* participants: does this mean then that they were not being creative? One thing we can rule out is that participants were simply copying the examples wholesale. In additional follow-up analyses of ideas generated in the *on-idle* and *on-demand* conditions, participants usually generated ideas that shared features with examples instead of simply copying them. Even in rare cases when participants submitted an idea that was almost identical to the examples, subsequent ideas were their own original ideas. We suspect that submitting ideas very similar to examples helped jolt their train of thoughts.

However, ruling out wholesale copying still leaves the question of whether ideas generated by solution transfer can be considered creative. We agree with other authors [Marsh et al., 1996, Purcell and Gero, 1996] that solution transfer *per se* does not mean that the resulting ideas are not creative (or were not produced by a creative process). Cognitive research strongly suggests that all idea generation is inevitably structured by prior knowledge [Ward, 1994], and studies of real-world creative behavior underscore the central importance of building on prior knowledge [Eckert and Stacey, 1998, Herring et al., 2009]. When this structuring and solution transfer leads to ideas that are novel and valuable, we say that the idea was "inspired by" or "built upon" the example(s) [Marsh et al., 1996, Herring et al., 2009]; in contrast, when the results are less desirable, we say that the designer was "fixated" by the examples [Linsey et al., 2010, Purcell and Gero, 1996]. Here, the fact that the *on-demand* participants mostly generated more novel ideas (and did not merely copy examples) suggests the former interpretation of the effects of examples is appropriate.

5.6.3 Further insights into the potential harm of examples

Our results also join prior work in highlighting the potential negative effects of examples. Here, we add the insight that at least some of the negative effects of examples may be due to when they are seen. Although participants in the *On-interval* condition generated ideas that were no less novel than those in the *On-demand* condition, they were the least productive (even less productive than people who saw no examples at all). One potential explanation—consistent with the SIAM model—might be that the examples were experienced as interruptions or distractions, rather than inspiration; much prior work has demonstrated that interruptions are detrimental to performance [Bailey et al., 2000]. Some authors have also suggested that interruptions and distractions can be especially detrimental when one is in a state of heightened focus and concentration on a creative task [Csikszentmihalyi, 1997]. While this effect might be caused by our choice of time interval, this result does demonstrate that it is possible to harm productivity with ill-timed example delivery. More in-depth examination of the effect of different length of time interval could shed some light on whether negative effects of fixed interval example delivery stem from poorly selected time intervals, or whether any fixed interval example delivery is likely to be suboptimal.

5.7 Conclusion

In this chapter we explored the question of how the impact of examples changes depending on when they are seen during ideation. We conducted an online experiment exploring two mechanisms for delivering examples at the right moment: a system that provides examples upon request and a system that proactively provides examples when a user is idle. Our results show that people benefit most from examples when they are prepared for it. Showing examples to people when they have been idle for a period of time helps people come up with more (but not necessarily better) ideas, while showing examples on-demand helps people come up with more novel ideas. In contrast, ill-timed example delivery might harm productivity, leading to fewer ideas.

Chapter 6

Summary View for Solution Synthesis

This chapter has adapted, updated, and rewritten content from a working menuscript in collaboration with Joel Chan, Steven P. Dow and Krzysztof Z. Gajos. All uses of “we”, “our”, and “us” in this chapter refer to coauthors of this work.

While open online innovation platforms promise great benefits from a large number of ideas in the divergent phase of the creative process, these platforms pose a challenge during the convergent phase. Once they collect enough ideas, someone has to summarize generated ideas and synthesize a few solutions to pursue. The task of solution synthesis usually falls to people who organize ideation challenges, hired experts or representatives of the communities that use the platforms. We refer to people who synthesize solutions from collected ideas “synthesizers”. In this phase, the synthesizers develop general knowledge of the ideas (main categories of ideas and their distribution), identify promising ideas including those rare gems, and craft solutions from what they learn from submitted ideas. However, the large number of ideas with different levels of detail and clarity makes these tasks difficult. Synthesizing these ideas for a few solutions typically involves looking through all ideas; a long and tiring process that biases them towards common solutions instead of rare and creative ones. In this chapter, we propose using a summary view that helps synthesizers learn about main

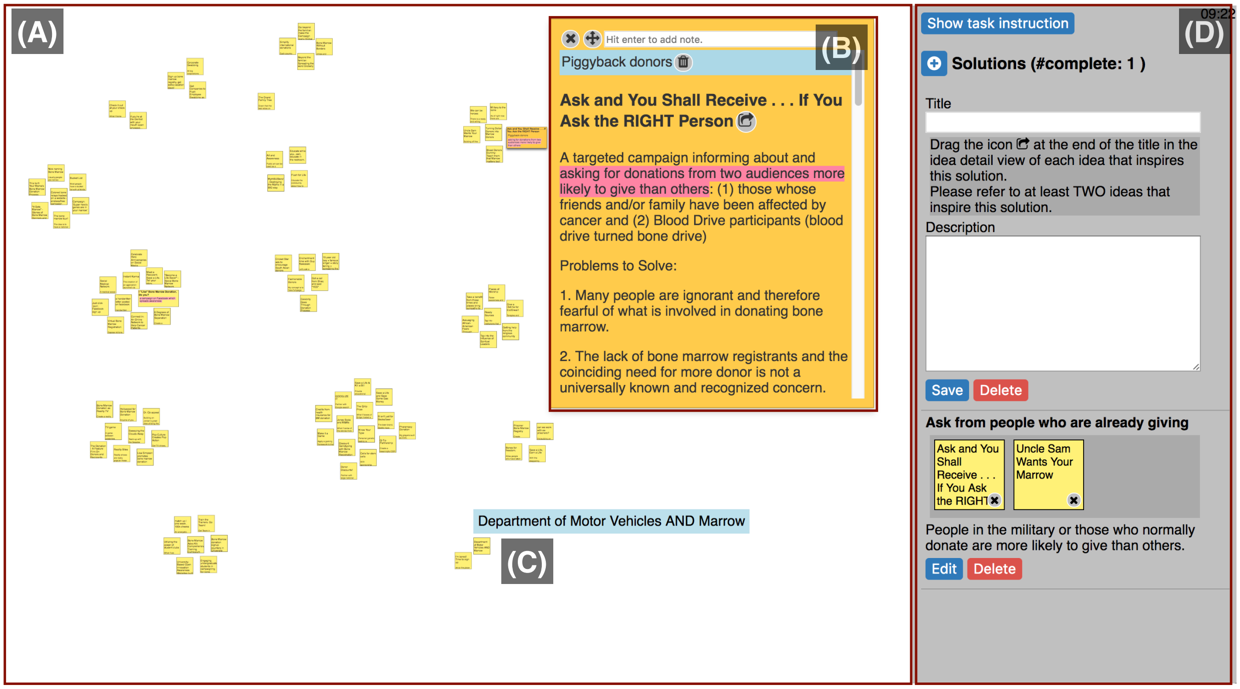


Figure 6.1: Screenshot of the synthesis interface. (A) A virtual whiteboard where seed ideas are positioned in group based on their similarities to one another. Users can drag an idea around to rearrange the layout. (B) Users can click on an idea to read the idea's full description. They can also make note about an idea (shown in blue boxes on the top of the detail pane) and highlight parts of an idea. (C) When users hover over an ideas, a tooltip with the idea's title pops up for a quick read. (D) Users write their solutions in this pane. They clicked on the add button to start writing a new solution. Each solution require a title, a description and at least two seed ideas. Users save a solution by clicking the Save button. The users can edit a saved solution by clicking on an Edit button. They can also delete a solution.

categories of ideas and spot rare and creative ideas. Our approach uses an *idea map*, as described in Chapter 4, to generate idea space and creates a summary view, a visualization that shows ideas in groups based on their similarities. This summary view (Figure 6.1) also presents small groups of ideas (rare ideas) as important as common ideas. However, the summary view might fixate the synthesizers on a single set of categories. We explore this tradeoff by conducting an experiment asking participants to synthesize solutions from sets of 87 ideas. Some participants were provided with a summary view manually generated by the researchers while some were provided a visualization of randomly positioned ideas. We found that participants with a summary view processed more rare ideas and integrated more rare ideas in their solutions but were fixated to the schema suggested by the summary view. There was no difference in the number of synthesized solutions. These results help inform the design of a future summary view that provides quick access to the general knowledge of the idea space and rare ideas while mitigating the fixation caused by the summary view.

6.1 Motivation and Contributions

Open innovation process does not end when all ideas are collected. The gathered ideas in their raw forms, albeit abundant in number, are not immediately usable. Some ideas are not complete solutions, lead to bad solutions or simply replicate many other ideas. The large number of ideas also makes it impractical to implement all of them. To extract value from the collective effort, synthesizers—usually experts or main stakeholders—evaluate the ideas, combine appropriate ideas together and generate a few polished solutions to pursue. We call this process "solution synthesis" and, for the rest of this chapter, refer to those responsible for this process as users.

Current solution synthesis involves looking through all ideas, comparing ideas against each other, evaluating ideas and synthesizing solutions from multiple ideas. This process requires a lot of time and effort from users. For example, Cambridge participatory budgeting 2016's idea synthesis took 60 representatives (Budget Delegates) 3 months to synthesize 20 solutions from 548 raw ideas[1]. The most time consuming part is understanding all ideas. Users can save time by comprehending a subset of ideas (e.g., the most popular ones or random ones). However, depending on the subset they select, they might overlook some promising rare ideas. With no knowledge of an overview of the idea space, the users have no way of identifying a

[1]http://pb.cambridgema.gov/pbcycle3

subset of ideas that allows them to gather important concepts while minimizing the time.

Existing platforms such as OpenIDEO[2] and MyStarbucksIdea.com[3] have used simple voting mechanisms to select ideas that are most popular. However, popular ideas do not always translate to solutions that are considered best by the stakeholders. For example, some ideas with high votes on OpenIDEO did not become winning ideas. It is therefore more useful for the synthesizers to get a holistic view of the idea space rather than narrow view on ideas with high popular votes. More importantly, these mechanisms might overlook rare ideas that are not seen by many and thus received fewer votes [Xu and Bailey, 2012]. A Budget Delegate from Cambridge participatory budgeting also told us in an informal interview that they ignored the popular votes of ideas and had to process all ideas anyway. To effectively synthesize solutions from a large set of ideas, the synthesizers need to be able to make senses of the emerging solution space and judiciously compare different possible solutions to one another.

One of the most time-consuming parts of solution synthesis for large-scale collective ideation is processing all ideas that are mostly mundane and redundant [Klein and Garcia, 2015, Bjelland and Wood, 2008]. In this context, an interpretative summary view that groups similar ideas together will save a user some time from repeatedly processing similar ideas. Further, prior work has suggested that a summary view that reveals schema of ideas improves a user's sensemaking [Russell et al., 2006, Fisher et al., 2012, Kittur et al., 2014], especially when the summary matches with the user's mental representation [Tversky et al., 2006]. The users can develop better understanding of the emerging solutions from the summary view and thus make informed decision about which solutions to pursue.

Prior work has explored aiding users in making sense of a large set of information by

[2]https://openideo.com/

[3]https://www.starbucks.com/coffeehouse/learn-more/my-starbucks-idea

providing a summary in different formats. Apolo provides users an interactive summary view that shows similar items grouped together based on the users' evolving mental model but is limited to network data such as paper citations [Chau et al., 2011]. Kittur et al. [2014] proposes a summary with attributes of items provided by previous users. Idea Spotter provides a summary of core parts of ideas marked by other users [Convertino et al., 2013]. IdeaGens uses a dashboard with a word cloud visualization of submitted ideas to summarize evolving solution space to support facilitating synchronous ideation [Chan et al., 2016]. Both Idea Spotter and IdeaGens did not give information about how ideas are grouped semantically and were limited to ideas expressed in text. Russell et al. [2006] and Gumienny et al. [2014] explore summary views that cluster similar items together visually. Grokker2 demonstrated the benefits of an interactive summary view that allows users to scan news articles quickly and move the around to reshape semantic clusters to fit their needs [Russell et al., 2006]. Qualitative findings from Gumienny et al. [2014] indicate that seeing how others cluster ideas and comparing them with one's own way of clustering can help provide different perspectives. These summary views however were genereated from inputs from a small group of users and automated methods instead of a crowd contribution.

Our approach is to leverage already derived idea map representation during idea generation to create a summary view that presents ideas in groups based on the ideas' similarities to one another. An idea map (Chapter 4) already has information about how ideas are related to each other and we can generate this summary view from an idea map by clustering similar ideas on the idea map together into groups. This summary view shows users an overview of main categories of the ideas. It also makes less common ideas as visually prominent as common ideas (Figure 6.1). The summary view suggests that rare ideas are worth inspecting as much as common ones and reduces the burden of inspecting a lot of similar ideas from the users. We therefore hypothesized that the users presented with an automated synthesis would be more likely to include rare ideas in their solutions than users working with no summary view. Further, because inspecting a summary view requires reviewing only a fraction of the ideas to make sense of each of the groups, we hypothesized that users working with the summary view would perceive lower task load than users reviewing raw ideas without any organization imposed on them.

However, there is a trade-off in this design. The summary view might fixate the users toward certain schemas or clustering of ideas [Barsalou, 1983, Nijstad and Stroebe, 2006]. Ideas are multi-faceted. There are usually multiple schemas or points of view to organize information [Barsalou, 1983, Gumienny et al., 2014, Chi et al., 1981]. By getting exposed to a single schema suggested by the summary view, the users might get fixated on that particular schema instead of trying to look at the solution space from different points of view. This is problematic because the users could miss some insights about the ideas that lead to good solutions.

We conducted an experiment to study this tradeoff. We asked 79 participants to synthesize solutions from ideas taken from a real ideation challenge. For each task, participants synthesized as many solutions as they could from 87 ideas using one of the two systems: with a summary view or with a visualization that positions ideas randomly instead of grouping them similar ones together. Our study measures how likely the participants would adopt rare ideas (ideas that have at most 2 ideas that share the same concept) to their solutions and whether participants with a summary view were fixated on the schema suggested by the visualization. Our results demonstrate that users with a summary view process and integrate rare ideas more than those without a summary view but also fixate more on the groups suggested by the summary view.

6.2 Experiment

We want to compare experience and behaviors of the users who were provided with a summary visualization of ideas to the users who were provided with a set of ideas with no summary view. The goal is to understand the potential trade-off of providing users with a summary visualization.

With a summary visualization, users do not have to create the schemas for categories of ideas by themselves. They would therefore have more time to synthesize new solutions. Our version of summary visualization also tells the users which ideas are rare or common. By featuring rare ideas in their own groups in a summary view, the users can identify these rare ideas easily. However, the summary visualization might fixate the users on the schema suggested by the summary when there are possibly other schemas that can help with synthesizing solutions.

Based on these arguments, we thus hypothesize:

H1: Users with a summary visualization have less task load and synthesize more solutions than users without it.

H2: Users with a summary visualization interact more with rare ideas as proposed by the visulization than users without it.

H3: Users with a summary visualization are fixated more on the categories suggested by the visualization than users without it.

6.2.1 Participants

We recruited 85 participants from Amazon Mechanical Turk (MTurk), an online micro-labor market. We limited recruitment to workers who resided in the U.S. and who had completed at least 1,000 HITs with greater than 95% approval rate. Participants were paid \$3.75 (\$9/hour) for their participation.

6.2.2 Task

Participants were asked to synthesize as many creative solutions as they could within 15 minutes. They were asked to synthesize the solutions from existing seed ideas that aimed to increase the number of bone-marrow donors. For each solution, a participant provided a title, a short description and a list of at least two seed ideas that inspired the solution.

6.2.3 Seed Ideas

We selected 87 seed ideas from 279 ideas submitted to an OpenIDEO's challenge on increasing the number of registered bone-marrow donors[4]. The quality of a summary view derived from an idea map depends on many factors such as the quantity and quality of human inputs and parameters of clustering algorithms. For this experiment, we decided to carefully control the quality of the summary view and grouped similar ideas manually. We selected seed ideas as follow. We read through all 279 ideas and, after filtering out ideas that were hard to understand without visual images or external links, clustered ideas into groups. From these groups, we selected 16 groups that represented most of the solution space without overlapping one another. We then further removed some ideas from some groups to create groups with fewer ideas . The list of seed ideas and their corresponded groups can be found in Appendix B.

We defined a rare idea as an idea that belongs to a group with at most 3 ideas. According to this threshold, there were 8 rare seed ideas making up 9.1% of all ideas.

[4]https://challenges.openideo.com/challenge/how-might-we-increase-the-number-of-bone-marrow-donors-to-help-save-more-lives

6.2.4 Procedure

Participants first read the description of the task and followed a tutorial that walked them over the interfaces they used to synthesize solutions. They answered a pre-task question that evaluated their self-efficacy in synthesizing solution: *"(On a scale of 1 being not at all confident and 7 being very confident) How confident are you that you can synthesize diverse and creative solutions from a large number of ideas generated by others?"*. Then, they read the description of the ideation challenge and spent 15 minutes synthesizing the solutions using the provided synthesis interface (Figure 6.1 or Figure 6.2). During these 15 minutes, participants wrote solutions based on seed ideas. They could hover over an idea to read its full title, click on an idea to open a window with its full description, write notes to idea, highlight text in an idea, and move an idea around the whiteboard. After they finished synthesizing solutions, participants answered questions about their experience.

At the beginning of the experiment, each participant was randomly assigned to one of the two conditions:

- *Summary*: Participants were initally presented with a whiteboard on which ideas were grouped together based on how similar they are to each other as shown in Figure Figure 6.1.
- *Random*: This is the baseline condition. Participants were initially presented with a whiteboard on which ideas were placed randomly (Figure 6.2).

6.2.5 Measures and Analysis

We conducted a between-subjects study with the two conditions as the sole factor on the following measures.

- *Number of valid solutions*. We counted the number of undeleted solutions with a title, a descrpition and at least two source ideas.

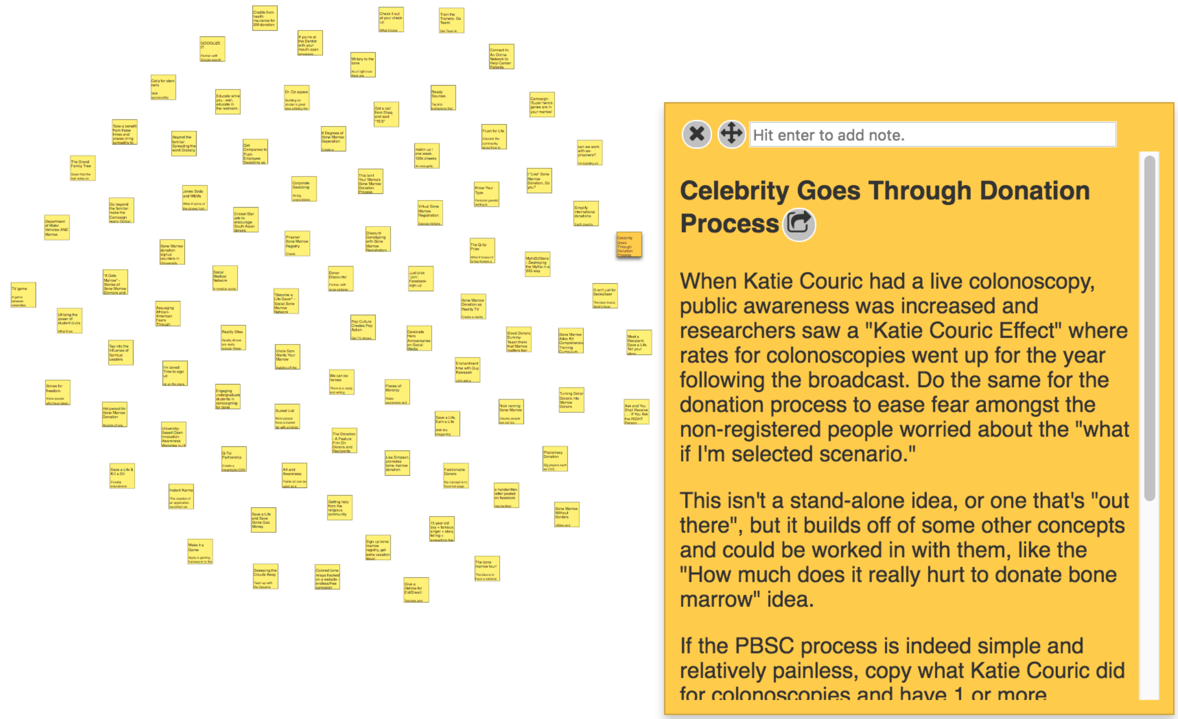


Figure 6.2: An example of the random intial positions of seed ideas on the whiteboard for the *Random* participants

- *Rare idea exploration.* We used the ratio of rare ideas over all ideas that participants hovered and clicked open and the ratio of rare ideas over all unique ideas that are integrated into valid solutions to measure their exposure to rare ideas.

- *Category fixation.* We used the ratio of valid solutions that cite seed ideas from more than one of the 16 groups that were presented in the *Summary* conditions over all solutions. The higher the ratio, the more likely participants synthesized ideas from more than one groups and signifies that participants were less fixated by the suggested grouping presented in the *Summary* condition.

We also collected participants' subjective response (reported on a 7-point Likert scale) to questions that related to their experience:

- *Self-efficacy* We compare the differences between pre-task and post-task self-efficacy in synthesizing solutions from a large number of seed ideas.

- *Perceived helpfulness of the initial positioning of seed ideas.* We asked participants to rate how helpful the initial positioning of seed ideas help them spot rare ideas and provide big picture (overview) of the ideas. We also asked how well the positioning match with their interpretation of the semantic similarities between ideas.
- *Perceived task load* We used the standard NASA Task Load Index (TLX) questions to measure workload perceived by the participants.

To reduce the probability of Type I error when performing multiple tests, we applied the Holm's sequentially-rejective Bonferroni procedure[Holm, 1979, Shaffer, 1995]. The procedure was applied separately to participants' subjective responses and separately to non-subjective performance measures (number of solutions, rare idea exploration and category fixation).

6.2.6 Adjustments to the data

We filtered out 4 participants who did not submit any valid solutions and 2 participants who submitted solutions that were not related to the the task.

We ended up with 79 participants: 41 in the *Random* condition and 38 in the *Summary* condition. 29 participants were female; 49 participants were male and 1 participant preferred not to identify themselves as either.

6.3 Results

We summarize the results for performance measures in Table 6.1 and the results of participants' subjective responses in Table 6.2. Table 6.3 synthesizes all results.

Measure (Hypothesis)	Grouped Mean (SD)	Random Mean (SD)	Raw p-value	Adjusted p-value
Number of valid solutions (H1)	3.26 (2.02)	2.98 (1.54)	0.4774	1.4322
Number of hovered ideas	54.03 (18.25)	55.49 (18.40)	0.7243	1.4486
Number of open ideas	19.29 (12.66)	19.39 (12.70)	0.9719	0.9719
Ratio of hovered rare ideas (H2)	0.0978 (0.0292)	0.0918 (0.0284)	0.3543	1.4172
Ratio of open rare ideas (H2)	0.1491 (0.1163)	0.0824 (0.0913)	0.0057	0.0285 *
Ratio of cited rare ideas (H2)	0.1721 (0.1852)	0.0767 (0.1040)	0.0052	0.0312 *
Ratio of solutions that cite ideas from different category (H3)	0.40 (0.43)	0.82 (0.29)	<.0001	<.0007 *

Table 6.1: Measures of participants' performance and interactions.

6.3.1 No substantial difference in the number of valid solutions

On average, the *Summary* participants synthesized 3.26 solutions (SD=2.02), while the *Random* participants syntehsized 2.98 solutions (SD=1.54). This difference is not significant ($F(1, 77) = 0.5097$, $p = 0.4774$). These results provide no support for H1.

6.3.2 No substantial difference in the number of hovered and clicked ideas

The *Summary* participants hovered, on average, over 54.03 ideas—62.10% of all seed ideas—(SD=18.25) to read the ideas' titles. The *Random* participants hovered over 55.49 ideas—63.78% of all seed ideas—(SD=18.40). There is no significant difference in the number of hovered ideas between the two conditions ($F(1, 77) = 0.1253$, $p = 0.7243$).

For deeper processing of ideas, participants could click on an idea to read its full description. On average, the *Summary* participants clicked open 19.29 ideas (SD=12.66) while the *Random* participants clicked open 19.39 ideas (SD =12.70). The difference of number of open ideas between conditions is not statistically significant ($F(1, 77) = 0.0012$, $p = 0.9719$).

6.3.3 Participants from both conditions hovered over equal ratio of rare ideas but the *Summary* participants clicked open and adopted rare ideas in higher ratio to their solutions

Out of all hovered ideas by the *Summary* participants, on average 9.78% (SD=2.92) were rare ideas. This percentage is slightly higher than that of the *Random* participants with 9.18% (SD=2.84). The difference of ratio of rare hovered ideas between the two conditions is not statistically significant ($F(1, 77) = 0.8685$, $p = 0.3543$).

In constrast, out of all *open* ideas by the *Summary* participants, on average 14.91% (SD=11.63) were rare ideas. The percentage is higher than that of the *Random* participants with 8.24% (SD=9.13). The difference of ratio of rare open ideas between the two conditions is statistically significant ($F(1, 77) = 8.1059$, $p = 0.0057$).

Likewise, out all ideas that the *Summary* participants cited, on average 17.21% (SD = 18.52) were rare ideas. The percentage is higher than that of the *Random* participants with 7.67%(SD=10.40). The difference of ratio of rare cited ideas between the two conditions is statistically significant ($F(1, 77) = 8.2728$, $p = 0.0052$).

These results provide support for H2.

6.3.4 The *Summary* participants fixated more on category suggested by the visualization

On average, the ratio of solutions that cite ideas from different groups over all solutions for participants in the *Summary* condition is 0.40 (SD=0.43), which is significantly lower than those of participants in the Random condition (0.82, SD=0.29). The difference between

Measure	Questions	Grouped Mean (SD)	Random Mean (SD)	Raw p-value	Adjusted p-value
Difference between pre-task and post-task self-efficacy	How confident are you that you can synthesize diverse and creative solutions from a large number of ideas generated by others? (Report increase of the post-task response from the pre-task response)	0.03 (1.91)	0.49 (1.57)	0.2426	1.4556
Perception of helpfulness of the initial positioning of seed ideas	Q1: How much of the big picture of ideas you got from this session?	5.71 (1.18)	5.20 (1.14)	0.0528	0.4224
	Q2: How helpful the system was in helping spotting rare ideas (ideas that have concepts that are shared by no or few other ideas)?	5.26 (1.64)	4.15 (1.86)	0.0061	0.0549
	Q3: How well does the initial positions of ideas on the whiteboard match with your interpretation of the semantic similarities between ideas?	5.50 (1.43)	3.51 (1.69)	<.0001	<0.001 *
Perceived workload	Q4: How mentally demanding was the task?	5.82 (1.14)	5.54 (1.25)	0.3026	1.513
	Q5: How physically demanding was the task?	2.05 (1.54)	2.12 (1.49)	0.8393	0.8393
	Q6: How hurried or rushed was the pace of the task?	5.32 (1.56)	4.68 (1.63)	0.0829	0.5803
	Q7: How successful were you in accomplishing what you were asked to do?	4.76 (1.24)	4.88 (1.25)	0.6830	1.366
	Q8: How hard did you have to work to accomplish your level of performance?	5.84 (1.03)	5.66 (1.17)	0.4635	1.854
	Q9: How insecure, discouraged, irritated, stressed, and annoyed were you?	3.42 (1.97)	3.19 (1.66)	0.5820	1.746

Table 6.2: Participants' subjective responses. The *Summary* participants found the initial positioning of ideas matched with their interpretation of the semantic similarities between ideas significantly more than those in the *Random* condition.

the two conditions is statistically significant ($F(1, 77) = 25.8180$, $p < .0001$). This means that participants in the *Random* conditions are more likely to propose solutions that got inspired by ideas from different groups suggested by the visualization seen by participants in the *Summary* condition. These results provide support for H3.

6.3.5 No substantial difference in the difference between pre-task and post-task self-efficacy

On average, the self-efficacy after the task increased by 0.03 (SD=1.91) for participants in the *Summary* and 0.49 (SD=1.57) for participants in the *Random* condition. There is no significant difference between the two conditions ($F(1, 77) = 1.3867$, $p = 0.2426$). See the adjusted p-value in Table 6.2.

6.3.6 Perceived helpfulness of the initial positioning of seed ideas

Question Q1 to Q3 in Table 6.2 measured participants' perceived helpfulness of the initial positions of seed ideas. We found no significant difference in perceived helpfulness in providing overview of the ideas across conditions. We found no significant different in perceived helpfulness in spotting rare ideas after applying the Holm's Bonferroni correction (adjusted $p = 0.0549$). However, participands in the *Summary* condition reported that the initial layout matched their semantic similarities mental model significantly more than the *Random* participants ($p < .0001$, adjusted $p < .001$). These results provide partial support to H2.

6.3.7 No substantial difference in perceived task load

Question Q4 to Q9 in Table 6.2 measured the participants' perceived task load while synthesizing ideas. We found no significant difference of perceived mental demand, physical demand, temporal demand, performance, effort and frustration. These results provide no support for H1.

6.4 Discussion

6.4.1 Number of synthesized solutions

We initially hypothesized that the *Summary* participants would synthesize more solutions than the *Random* participants. The former did not have to construct the schema of the solution space from scratch so we had expected the *Summary* participants to have more time to focus on synthesizing ideas. However, our results show no differences in the number of synthesized solutions across conditions. Further inspection on the length of written solutions also show no differences across conditions on how much the participants wrote and how many

Hypothesis	Measure	Hypothesis supported
H1	No substantial difference in the number of submitted solutions	-
	No substantial difference in perceived task load	-
H2	No substantial difference in ratio of hovered rare ideas	-
	The Summary participants inspected higher ratio of rare ideas than the Random participants	Yes
	The Summary participants cited higher ratio of rare ideas than the Random participants	Yes
	No substantial difference in perceived helpfulness in providing an overview of the ideas	-
	No substantial difference in perceived helpfulness in spotting rare ideas	-
	The Summary participants reported the initial idea layout matched their mental model more than the Random participants	Yes
H3	The Summary participants submitted lower ratio of solutions that cite ideas from different categories	Yes

Table 6.3: Summary of performance measures and subjective responses for each hypothesis

idea sources they cited per solution.

One explanation is that participants in both conditions were equally pressured by the time limit and had to distribute the time accordingly. This corresponded with the survey results where participants from both conditions reported that they were equally rushed. We set the time limit to 15 minutes so that participants felt slightly rushed even though they only had 87 ideas to explore instead of hundreds. Participants had to make a trade-off between exploration, deciding on the solutions to pursue and writing the solutions. An experiment with longer time limit might reveal a more informative picture of how participants balance these activities.

6.4.2 Discovering rare ideas with a summary view

The results demonstrated that a summary view that shows ideas grouped semantically helps users spot rare ideas that they might overlook otherwise. The *Summary* participants clicked open rare ideas more than the *Random* participants even though both initially hovered the same ratio of rare ideas. These results suggested that the *Random* participants were not aware that an idea was rare when they hovered over it and thus did not click on them to process them further. By presenting rare ideas in their own clusters, our summary view makes it easier for users to spot uncommon ideas which can lead to creative solutions.

6.4.3 Fixating on categories suggested by the summary view

We asked participants to cite at least two ideas for each solutions. Citing ideas from different groups implies that the participants find commonality (semantic similarities) between the ideas that were not grouped together in the summary view. Our results show that the users with a summary view synthesized fewer solutions with cross-group ideas as predicted by prior work on cognitive fixation [Nijstad and Stroebe, 2006, Kohn and Smith, 2011]. Seeing a summary view that shows only one point of view can prevent participants from synthesizing solutions that could have been derived from an alternative view. For example, a participant submitted a solution "Mandatory Donation" that proposed making the donation mandatory for people who are already giving, such as military officers and blood donors, but the participant did not propose a solution with another aspect of the seed ideas in the groups that proposed demystifying the donation process.

One approach that might mitigate the schema fixation effect is to expose users to different schemas. Instead of showing just only one way to categorize ideas, a summary view can show multiple ways to group ideas semantically. Building on our approach to infer semantic relationships in Chapter 5, we can leverage different ways IdeaHound users group ideas to generate multiple idea maps instead of simply aggregating them to generate a single idea map. For example, we can apply a machine learning algorithm to identify different types of users based on how they cluster ideas on the whiteboard and then generate an idea map for each type [Kairam and Heer, 2016] or derive a latent factor model to learn similarity functions of ideas for a user population [Yue et al., 2014]. The latter approach also supports personalized clustering inferences where the algorithm tries to predict how the users will cluster ideas based on how they cluster a small subset of ideas. This approach could help users to gradually develop a schema that fits with their mental models while also offering alternative perspectives.

We also note that fixation on categories might not necessarily be harmful, especially during the convergent phase of the ideation process. Focusing on a few categories presented could ease the decision making process and give the synthesizers more time to prototype and test the solutions they generate. Future work could explore the benefits and setbacks of category fixation during the convergent phase.

6.5 Conclusion

In this chapter, we explored a summary view interface that helps synthesizers synthesize solutions from a large set of raw ideas. We presented an experiment that studied the trade-offs of the proposed summary view. Our results demonstrated that the summary view helps users find rare ideas but it fixates users on certain schemas. We discussed this trade-off and proposed alternative solutions to defixate the users while still retaining the benefit of the presented summary view.

Chapter 7

Conclusion and Future Directions

This dissertation addresses challenges in large-scale collective ideation. It argues that an intelligent system that understands the emerging solution space of ideas can improve people's creative output in both the convergent phase and the divergent phase of ideation. I support the dissertation's thesis by presenting empirical studies of creativity enhancing interventions, introducing a computational model that helps intelligent systems understand the emerging solution space, and developing a system IdeaHound to demonstrate how an intelligent system can use the computational model to support idea generation. I summarize the contributions of this dissertation and discuss future directions in the following sections.

7.1 Contributions

This dissertation makes the following contributions.

- **Knowledge about creativity enhancing interventions**: I presented findings that corroborate and complement theories and results from existing creativity research. My empirical studies were motivated by system interventions that could enhance the creative output of users. In Chapter 4, I demonstrated that showing a diverse set of examples selected from a computational model yields more diverse ideas as predicted by existing cognitive models and research. In Chapter 5, I provided evidence that people benefit best from examples when they are prepared to receive them. This result was predicted by the SIAM model but had not been empirically tested.

- **Domain-independent computational model of ideas**: Existing large-scale online ideation platforms lack ways to help contributors to deliberately explore ideas of others for inspirations. Prior work has suggested that deliberate exploration interactions such as looking at a set of diverse ideas, looking at a set of similar ideas or getting an overview of the solution space can improve creative output. To enable such interactions, I introduced a computational model that encodes similarities between ideas. I also demonstrated how to derive this model at scale and how to use the model to enable creativity enhancing interactions during both the divergent phase (Chapter 3 and Chapter 4) and the convergent phase (Chapter 6) of ideation. The application of such a model to support creative ideation is a novel and enduring contribution.

- **Demonstration of integrated crowdsourcing**: I introduced the concept of "integrated crowdsourcing", an approach that integrates the potentially tedious secondary task with the more intrinsically-motivated primary task. Unlike traditional micro-task approach, the integrated crowdsourcing approach requires neither significant extra human resources nor extrinsic incentives to motivate people to do the micro-tasks. I applied integrated crowdsourcing to support collective ideation and illustrated the approach's effectiveness in collecting information to support collective activities in one domain. The general approach of integrated crowdsourcing I proposed in Chapter 4 can be extended to gather other information about ideas and be applied to other domains. Integrated crowdsourcing is especially useful in social computing domains that rely on intrinsically motivated volunteers due to lack of funding or nature of the desired information.

7.2 Future Directions

My dissertation work addresses a subset of the many challenges of collective ideation. In this section, I enumerate future directions to explore.

7.2.1 Improve the computational models of ideas

Idea map has limitations. The model was designed to be light-weighted to minimize the work required from humans. The result is a model that has minimal information to help people navigate the idea space. However, it also limits the kind of creativity enhancing interventions the model enables. For example, an idea map cannot explain why certain ideas are similar or different or how to combine multiple ideas together. Future work could explore a model with more useful information such as category labels or values of important dimensions of the idea space. The challenge to doing so is identifying ways to extract such information efficiently for a large number of ideas.

Another way to extend the impact of the idea map model is to reduce the amount of human effort in computing this model. The approaches that generate an idea map described in this dissertation rely solely on human judgments because ideas can come in many formats. Still, a lot of ideas are expressed in text. Advances in Natural Language Processing research have now enabled efficient methods to understand short text snippets [Mikolov et al., 2013, Pennington et al., 2014]. Exploring integrating automated results from such methods is a promising future direction. For example, an automated method can provide seed information about relationships between ideas that people could correct or people could provide information that the automated method falls short [Chang et al., 2016].

7.2.2 Personalized inspirations

Many factors, such as timing and a person's cognitive state, determine how example ideas affect their idea generation. The findings from this dissertation work and a follow-up study in a collaboration with Chan [Chan et al., 2017] point toward the benefits of presenting personalized examples that would be most helpful to people at a specific point in time. Future work could explore modeling a user's cognitive states: whether they are on a roll or stuck and the most recent topic they have explored. An intelligent system with such a model could infer the optimal time to present personalized inspirations to an individual.

7.2.3 Coordinate a community effort

This dissertation explored creativity enhancing interventions of a single contributor in both divergent and convergent phases. Coordinating efforts from multiple contributors could further improve the performance of the community as a whole. For example, a community can use an idea map as a guideline to coordinate effort of contributors who explore different parts of the emerging solution space. Some contributors might develop expertise to deeply explore specific parts of the solution space. When these contributors work together, communities benefit from both breadth and depth of collective solution space exploration.

For the convergent phase, I envision that synthesizers could work together or with the rest of the community to synthesize well-thought out and comprehensive solutions. For example, multiple representatives can exchange their points of view to help each other develop comprehensive perspectives of the solution space. Other community members can help organize and filter ideas and solutions in the background while the representatives synthesize the solutions. An intelligent system can act as a bridge between different contributors and coordinate their effort based on their roles, capabilities, experiences and preferences.

7.2.4 Collective ideation for complex problems

The summary view proposed in Chapter 6 assumes that ideas do not depend on each other and are of the same level of granularity. However, many real-life complex problems require solving interdependent sub-problems and balancing a range of constraints and preferences that emerge during the ideation process. For example, a mass transportation system that comprises multiple modes of transportation would require designing and integrating solutions for complex sub-problems, such as traffic flow, coordination system and maintenance support. These sub-problems are linked to each other in complex and mutually dependent ways. They could benefit from large-scale collective ideation, where many explore and refine the solutions for sub-problems in parallel. However, most proposed ideas are generated separately. We could provide contributors with a way to coordinate their effort to ensure that most contributions will eventually lead to successful integrated solutions.

7.3 Collective ideation in the real world

Large creative online platforms could transform the way our society innovates. They make it possible for anyone to contribute to the problem that they care about, democratize and increase transparency of important decision-making processes, and promise to jumpstart innovation process by leveraging the immense diversity of perspectives and experiences of contributors. Although some might currently view collective ideation as a marketing gimmick, it has potential to solve real large problems that affect a large group of people. We can turn this potential into reality by addressing the challenges of scaling faced by current open innovation platforms.

My research addresses some challenges of large-scale collaborative innovation and proposes technical approaches for a more effective collaboration. It is a small step toward supporting real-world collaborative innovation where many other factors can lead to successful or unsuccessful outcomes. My dissertation work contributes to improving the current state of crowd innovation, so that we can effectively harvest people's collective effort to make the world a better place.

Bibliography

Marine Agogué, Akin Kazakçi, Armand Hatchuel, Pascal Masson, Benoit Weil, Nicolas Poirel, and Mathieu Cassotti. The impact of type of examples on originality: Explaining fixation and stimulation effects. *The Journal of Creative Behavior*, 48(1):1–12, 2013.

Paul André, Aniket Kittur, and Steven P. Dow. Crowd synthesis: Extracting categories and clusters from complex data. In *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing*, CSCW '14, pages 989–998, New York, NY, USA, 2014. ACM. ISBN 978-1-4503-2540-0. doi: 10.1145/2531602.2531653. URL http://doi.acm.org/10.1145/2531602.2531653.

Brian P Bailey and Shamsi T Iqbal. Understanding changes in mental workload during execution of goal-directed tasks and its application for interruption management. *ACM TOCHI*, 14(4):21, 2008.

Brian P Bailey, Joseph A Konstan, and John V Carlis. Measuring the effects of interruptions on task performance in the user interface. In *Proc. IEEE SMC*, volume 2, pages 757–762. IEEE, 2000.

Michelle Q Wang Baldonado and Terry Winograd. Sensemaker: an information-exploration interface supporting the contextual evolution of a user's interests. In *Proceedings of the ACM SIGCHI Conference on Human factors in computing systems*, pages 11–18. ACM, 1997.

Linden J Ball, Thomas C Ormerod, and Nicola J Morley. Spontaneous analogising in engineering design: a comparative analysis of experts and novices. *Design Studies*, 25(5): 495–508, 2004.

Lawrence W Barsalou. Ad hoc categories. *Memory & cognition*, 11(3):211–227, 1983.

Jonali Baruah and Paul B. Paulus. Category assignment and relatedness in the group ideation process. *Journal of Experimental Social Psychology*, 47(6):1070–1077, 2011. ISSN 0022-1031.

Christopher Bearman, Linden J Ball, and Thomas C Ormerod. An exploration of real-world analogical problem solving in novices. In *Prog. CogSciâĂŹ02*, 2002.

Merim Bilalić, Peter McLeod, and Fernand Gobet. Why good thoughts block better ones: The mechanism of the pernicious einstellung (set) effect. *Cognition*, 108(3):652–661, 2008.

Osvald M Bjelland and Robert Chapman Wood. An inside view of ibm's' innovation jam'. *MIT Sloan management review*, 50(1):32, 2008.

David M Blei. Probabilistic topic models. *Communications of the ACM*, 55(4):77–84, 2012.

Kevin J Boudreau and Karim R Lakhani. âĂIJopenâĂİ disclosure of innovations, incentives and follow-on reuse: Theory on processes of cumulative innovation and a field experiment in computational biology. *Research Policy*, 44(1):4–19, 2015.

Danielle Bragg, Kyle Rector, and Richard E. Ladner. A user-powered american sign language dictionary. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*, CSCW '15, pages 1837–1848, New York, NY, USA, 2015. ACM. ISBN 978-1-4503-2922-4. doi: 10.1145/2675133.2675226. URL http://doi.acm.org/10.1145/2675133.2675226.

Bill Buxton. *Sketching User Experiences: Getting the Design Right and the Right Design (Interactive Technologies)*. Morgan Kaufmann, April 2007. ISBN 0123740371. URL http://www.amazon.com/exec/obidos/redirect?tag=citeulike07-20&path=ASIN/0123740371.

Joel Chan and Christian Schunn. The impact of analogies on creative concept generation: Lessons from an in vivo study in engineering design. *Cognitive Science*, 39(1):126–155, 2015. doi: 10.1111/cogs.12127.

Joel Chan, Katherine Fu, Christian Schunn, Jonathan Cagan, Kristin Wood, and Kenneth Kotovsky. On the benefits and pitfalls of analogies for innovative design: Ideation performance based on analogical distance, commonness, and modality of examples. *Journal of mechanical design*, 133(8):081004, 2011.

Joel Chan, Steven Dow, and Christian Schunn. Conceptual distance matters when building on others' ideas in crowd-collaborative innovation platforms. In *Proc. CSCWâĂŹ14*, pages 141–144. ACM, 2014.

Joel Chan, Steven Dang, and Steven P Dow. Improving crowd innovation with expert facilitation. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*, pages 1223–1235. ACM, 2016.

Joel Chan, Pao Siangliulue, Denisa Qori McDonald, Ruixue Liu, Reza Moradinezhad, Safa Aman, Erin Solovey, Krzysztof Z. Gajos, and Steven P. Dow. Semantically far inspirations considered harmful? accounting for cognitive states in collaborative ideation. In *Proceedings of the 2017 ACM SIGCHI Conference on Creativity and Cognition*, C&C '17, 2017.

Jonathan Chang, Sean Gerrish, Chong Wang, Jordan L Boyd-Graber, and David M Blei. Reading tea leaves: How humans interpret topic models. In *Advances in neural information processing systems*, pages 288–296, 2009.

Joseph Chee Chang, Aniket Kittur, and Nathan Hahn. Alloy: Clustering with crowds and computation. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pages 3180–3191. ACM, 2016.

Duen Horng Chau, Aniket Kittur, Jason I Hong, and Christos Faloutsos. Apolo: making sense of large network data by combining rich user interaction and machine learning. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 167–176. ACM, 2011.

Siddhartha Chaudhuri, Evangelos Kalogerakis, Stephen Giguere, and Thomas Funkhouser. Attribit: content creation with semantic attributes. In *Proceedings of the 26th annual ACM symposium on User interface software and technology*, pages 193–202. ACM, 2013.

Michelene TH Chi, Paul J Feltovich, and Robert Glaser. Categorization and representation of physics problems by experts and novices. *Cognitive science*, 5(2):121–152, 1981.

Lydia B Chilton, Greg Little, Darren Edge, Daniel S Weld, and James A Landay. Cascade: Crowdsourcing taxonomy creation. In *Proceedings of the 2013 ACM annual conference on Human factors in computing systems*, pages 1999–2008. ACM, 2013.

Lydia B. Chilton, Juho Kim, Paul André, Felicia Cordeiro, James A. Landay, Daniel S. Weld, Steven P. Dow, Robert C. Miller, and Haoqi Zhang. Frenzy: Collaborative data organization for creating conference sessions. In *Proceedings of the 32Nd Annual ACM Conference on Human Factors in Computing Systems*, CHI '14, pages 1255–1264, New York, NY, USA, 2014. ACM. ISBN 978-1-4503-2473-1. doi: 10.1145/2556288.2557375. URL http://doi.acm.org/10.1145/2556288.2557375.

Evangelia G Chrysikou and Robert W Weisberg. Following the wrong footsteps: fixation effects of pictorial examples in a design problem-solving task. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 31(5):1134, 2005.

Jason Chuang, Daniel Ramage, Christopher Manning, and Jeffrey Heer. Interpretation and trust: Designing model-driven visualizations for text analysis. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 443–452. ACM, 2012.

Jacob Cohen. *Statistical power analysis for the behavioral sciences (rev.* Lawrence Erlbaum Associates, Inc, 1977.

Gregorio Convertino, Ágnes Sándor, and Marcos Baez. Idea spotter and comment interpreter: Sensemaking tools for idea management systems. In *ACM Communities and Technologies Workshop: Large-Scale Idea Management and Deliberation Systems Workshop*, 2013.

William H. Cooper, R. Brent Gallupe, Sandra Pollard, and Jana Cadsby. Some liberating effects of anonymous electronic brainstorming. *Small Group Research*, 29(2):147–178, 1998.

Mihaly Csikszentmihalyi. *Flow and the Psychology of Discovery and Invention*. HarperPerennial, New York, NY, 1997.

Darren W Dahl and Page Moreau. The influence and value of analogical thinking during new product ideation. *Journal of Marketing Research*, 39(1):47–60, 2002.

A R Dennis and J S Valacich. Computer brainstorms: More heads are better than one. *Journal of Applied Psychology*, 78(4):531, 1993.

Michael Diehl and Wolfgang Stroebe. Productivity loss in brainstorming groups: Toward the solution of a riddle. *Journal of Personality and Social Psychology*, 53(3):497–509, 9 1987. ISSN 0022-3514.

Alex Doboli, Anurag Umbarkar, Varun Subramanian, and Simona Doboli. Two experimental studies on creative concept combinations in modular design of electronic embedded systems. *Design Studies*, 35(1):80–109, 1 2014. ISSN 0142694X. doi: 10.1016/j.destud.2013.10.002.

Steven Dow, Alana Glassco, Jonathan Kass, Melissa Schwarz, Daniel Schwartz, and Scott Klemmer. Parallel prototyping leads to better design results, more divergence, and increased self-efficacy. *Transactions on Computer-Human Interaction (TOCHI*, 17(4), 2010.

Steven Dow, Julie Fortuna, Dan Schwartz, Beth Altringer, Daniel Schwartz, and Scott Klemmer. Prototyping dynamics: Sharing multiple designs improves exploration, group rapport, and results. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '11, pages 2807–2816, New York, NY, USA, 2011. ACM. ISBN 978-1-4503-0228-9. doi: 10.1145/1978942.1979359. URL http://doi.acm.org/10.1145/1978942.1979359.

Claudia Eckert and Martin Stacey. Fortune favours only the prepared mind: Why sources of inspiration are essential for continuing creativity. *Creativity and Innovation Management*, 7(1):9–16, 1998.

Ronald A Finke, Thomas B Ward, and Steven M Smith. *Creative cognition: Theory, research, and applications*. MIT press Cambridge, MA, 1992.

Kristie Fisher, Scott Counts, and Aniket Kittur. Distributed sensemaking: improving sensemaking by leveraging the efforts of previous users. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 247–256. ACM, 2012.

Ronald S Friedman, Ayelet Fishbach, Jens Förster, and Lioba Werth. Attentional priming effects on creativity. *Creativity Research Journal*, 15(2-3):277–286, 2003.

R B Gallupe, A R Dennis, W H Cooper, J S Valacich, L M Bastianutti, and J F Nunamaker. Electronic brainstorming and group size. *Academy of Management Journal*, 35(2):350–369, 1992.

R Brent Gallupe, Lana M Bastianutti, and William H Cooper. Unblocking brainstorms. *Journal of Applied Psychology*, 76(1):137, 1991.

David Jason Gerber, Shih-Hsin Eve Lin, Bei Penny Pan, and Aslihan Senel Solmaz. Design optioneering: multi-disciplinary design optimization through parameterization, domain integration and automation of a genetic algorithm. In *Proceedings of the 2012 Symposium on Simulation for Architecture and Urban Design*, page 11. Society for Computer Simulation International, 2012.

Tim P German and H Clark Barrett. Functional fixedness in a technologically sparse culture. *Psychological Science*, 16(1):1–5, 2005.

Olga Goldenberg, James R. Larson, and Jennifer Wiley. Goal instructions, response format, and idea generation in groups. *Small Group Research*, 44(3):227–256, 2013. doi: 10.1177/1046496413486701. URL http://sgr.sagepub.com/content/44/3/227.abstract.

Robert Goldstone. An efficient method for obtaining similarity data. *Behavior Research Methods, Instruments, & Computers*, 26(4):381–386, 1994.

Ryan G. Gomes, Peter Welinder, Andreas Krause, and Pietro Perona. Crowdclustering. In J. Shawe-Taylor, R.S. Zemel, P.L. Bartlett, F. Pereira, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems 24*, pages 558–566. Curran Associates, Inc., 2011. URL http://papers.nips.cc/paper/4187-crowdclustering.pdf.

Joy Paul Guilford. The nature of human intelligence. 1967.

Raja Gumienny, Steven P Dow, and Christoph Meinel. Supporting the synthesis of information in design teams. In *Proceedings of the 2014 conference on Designing interactive systems*, pages 463–472. ACM, 2014.

Hannes Heikinheimo and A Ukkonen. The Crowd-Median Algorithm. *First AAAI Conference on Human . . .* , pages 69–77, 2013. URL http://www.aaai.org/ocs/index.php/HCOMP/HCOMP13/paper/view/7513.

Scarlett R Herring, Chia-Chen Chang, Jesse Krantzler, and Brian P Bailey. Getting inspired!: understanding how and why examples are used in creative design practice. In *Proc. CHIâĂŹ09*, pages 87–96. ACM, 2009.

Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian journal of statistics*, pages 65–70, 1979.

Michael C Hout, Stephen D Goldinger, and Ryan W Ferguson. The versatility of spam: A fast, efficient, spatial method of data collection for multidimensional scaling. *Journal of Experimental Psychology: General*, 142(1):256, 2013.

Cheng-Zhi Anna Huang, David Duvenaud, Kenneth C. Arnold, Brenton Partridge, Josiah W. Oberholtzer, and Krzysztof Z. Gajos. Active learning of intuitive control knobs for synthesizers using gaussian processes. In *Proceedings of the 19th International Conference on Intelligent User Interfaces*, IUI '14, pages 115–124, New York, NY, USA, 2014. ACM. ISBN 978-1-4503-2184-6. doi: 10.1145/2557500.2557544. URL http://doi.acm.org/10.1145/2557500.2557544.

Cheng-Zhi Anna Huang, David Duvenaud, and Krzysztof Z. Gajos. Chordripple: Recommending chords to help novice composers go beyond the ordinary. In *Proceedings of the 21st International Conference on Intelligent User Interfaces*, IUI '16, pages 241–250, New York, NY, USA, 2016. ACM. ISBN 978-1-4503-4137-0. doi: 10.1145/2856767.2856792. URL http://doi.acm.org/10.1145/2856767.2856792.

David G Jansson and Steven M Smith. Design fixation. *Design Studies*, 12(1):3–11, 1991a.

David G Jansson and Steven M Smith. Design fixation. *Design studies*, 12(1):3–11, 1991b.

Elahe Javadi. *The impact of user interface design on idea integration in electronic brainstorming: an attention-based view*. PhD thesis, University of Illinois at Urbana-Champaign, 2012.

Sanjay Kairam and Jeffrey Heer. Parting crowds: Characterizing divergent interpretations in crowdsourced annotation tasks. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*, pages 1637–1648. ACM, 2016.

Manu Kapur. Productive failure. *Cognition and Instruction*, 26(3):379–424, 2008.

Aniket Kittur, Bryant Lee, and Robert E. Kraut. Coordination in Collective Intelligence: The Role of Team Structure and Task Interdependence. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '09, pages 1495–1504, New York, NY, USA, 2009. ACM. ISBN 978-1-60558-246-7. doi: 10.1145/1518701.1518928. URL http://doi.acm.org/10.1145/1518701.1518928.

Aniket Kittur, Andrew M Peters, Abdigani Diriye, and Michael Bove. Standing on the schemas of giants: socially augmented information foraging. In *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*, pages 999–1010. ACM, 2014.

Mark Klein and Ana Cristina Bicharra Garcia. High-speed idea filtering with the bag of lemons. *Decision Support Systems*, 78:39–50, 2015.

Nicholas W Kohn and Steven M Smith. Collaborative fixation: Effects of others' ideas on brainstorming. *Applied Cognitive Psychology*, 25(3):359–371, 2011.

Steven Komarov and Krzysztof Z. Gajos. Organic peer assessment. In *Proceedings of the CHI 2014 Learning Innovation at Scale workshop*, 2014.

Robert E Kraut, Paul Resnick, Sara Kiesler, Moira Burke, Yan Chen, Niki Kittur, Joseph Konstan, Yuqing Ren, and John Riedl. *Building Successful Online Communities*. Evidence-Based Social Design. MIT Press, March 2012.

Chinmay Kulkarni, Steven P Dow, and Scott R Klemmer. Early and repeated exposure to examples improves creative work. In *Design Thinking Research*, pages 49–62. Springer, 2014.

Ranjitha Kumar, Arvind Satyanarayan, Cesar Torres, Maxine Lim, Salman Ahmad, Scott R Klemmer, and Jerry O Talton. Webzeitgeist: Design mining the web. In *Proc. CHIâĂŹ13*, pages 3083–3092. ACM, 2013.

Paul Laseau. *Graphic thinking for architects and designers*. John Wiley & Sons, 2001.

Edith Law and Luis Von Ahn. Input-agreement: a new mechanism for collecting data using human computation games. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 1197–1206. ACM, 2009.

Brian Lee, Savil Srivastava, Ranjitha Kumar, Ronen Brafman, and Scott R Klemmer. Designing with interactive example galleries. In *Proc. CHIâĂŹ10*, pages 2257–2266. ACM, 2010.

Karen Leggett Dugosh and Paul B Paulus. Cognitive and social comparison processes in brainstorming. *Journal of Experimental Social Psychology*, 41(3):313–320, May 2005.

Sheena Lewis, Mira Dontcheva, and Elizabeth Gerber. Affective computational priming and creativity. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 735–744. ACM, 2011.

JS Linsey, I Tseng, K Fu, J Cagan, KL Wood, and C Schunn. A study of design fixation, its mitigation and perception in engineering design faculty. *Journal of Mechanical Design*, 132 (4):041003, 2010.

Lan Luo and Olivier Toubia. Improving online idea generation platforms and customizing the task structure on the basis of consumers' domain-specific knowledge. *Journal of Marketing*, 79(5):100–114, 2015.

Norman RF Maier. Reasoning in humans. ii. the solution of a problem and its appearance in consciousness. *Journal of comparative Psychology*, 12(2):181–194, 1931.

A. Majchrzak and A. Malhotra. Towards an information systems perspective and research agenda on crowdsourcing for innovation. *The Journal of Strategic Information Systems*, 22(4):257–268, December 2013. ISSN 0963-8687. doi: 10.1016/j.jsis.2013.07.004. URL http://www.sciencedirect.com/science/article/pii/S0963868713000528.

Joe Marks, Brad Andalman, Paul A Beardsley, William Freeman, Sarah Gibson, Jessica Hodgins, Thomas Kang, Brian Mirtich, Hanspeter Pfister, Wheeler Ruml, et al. Design galleries: A general approach to setting parameters for computer graphics and animation. In *Proceedings of the 24th annual conference on Computer graphics and interactive techniques*, pages 389–400. ACM Press/Addison-Wesley Publishing Co., 1997.

Richard L Marsh, Joshua D Landau, and Jason L Hicks. How examples may (and may not) constrain creativity. *Memory & Cognition*, 24(5):669–680, 1996.

Richard L Marsh, Joshua D Landau, and Jason L Hicks. Contributions of inadequate source monitoring to unconscious plagiarism during idea generation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 23(4):886, 1997.

Philip Mendels, Joep Frens, and Kees Overbeeke. Freed: a system for creating multiple views of a digital collection during the design process. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 1481–1490. ACM, 2011.

Nicolas Michinov and Corine Primois. Improving productivity and creativity in online groups through social comparison process: New evidence for asynchronous electronic brainstorming. *Computers in Human Behavior*, 21(1):11–28, 1 2005. ISSN 07475632. doi: 10.1016/j.chb.2004.02.004.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013.

Jarrod Moss, Kenneth Kotovsky, and Jonathan Cagan. The influence of open goals on the acquisition of problem-relevant information. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 33(5):876, 2007.

Murli Nagasundaram and Alan R Dennis. When a group is not a group the cognitive foundation of group idea generation. *Small Group Research*, 24(4):463–489, 1993.

Jeffrey V. Nickerson and Lixiu Lisa Yu. Going Meta: Design Space and Evaluation Space in Software Design. In AndrÃľ van der Hoek and Marian Petre, editors, *Software Designers in Action: A Human-Centric Look at Design Work*, pages 323–344. Chapman and Hall/CRC, Rochester, NY, September 2012. URL http://papers.ssrn.com/abstract=2147158.

Jeffrey V Nickerson, James E Corter, Barbara Tversky, Doris Zahner, and Yun Jin Rho. The spatial nature of thought: Understanding information systems design through diagrams. In *in Boland, R., Limayem, M., and Pentland B.,(eds), Proceedings of the 29th International Conference on Information Systems*. Citeseer, 2008.

B A Nijstad, W Stroebe, and H F M Lodewijkx. Cognitive stimulation and interference in groups: Exposure effects in an idea generation task. *Journal of Experimental Social Psychology*, 38(6):535–544, 2002.

Bernard A Nijstad and Wolfgang Stroebe. How the group affects the mind: A cognitive model of idea generation in groups. *Personality and social psychology review*, 10(3):186–213, 2006.

Bernard A. Nijstad, Carsten K. W. De Dreu, Eric F. Rietzschel, and Matthijs Baas. The dual pathway to creativity model: Creative ideation as a function of flexibility and persistence. *European Review of Social Psychology*, 21:34–77, 2010a. ISSN 1046-3283. doi: 10.1080/10463281003765323.

Bernard A Nijstad, Carsten KW De Dreu, Eric F Rietzschel, and Matthijs Baas. The dual pathway to creativity model: Creative ideation as a function of flexibility and persistence. *European Review of Social Psychology*, 21(1):34–77, 2010b.

Peter O'Donovan, Jānis Lībeks, Aseem Agarwala, and Aaron Hertzmann. Exploratory font selection using crowdsourced attributes. *ACM Transactions on Graphics (TOG)*, 33(4):92, 2014.

Peter O'Donovan, Aseem Agarwala, and Aaron Hertzmann. Designscape: Design with interactive layout suggestions. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, pages 1221–1224. ACM, 2015.

SJ Parnes. Effects of extended effort in creative problem solving. *Journal of Educational psychology*, 52(3):117–122, 1961. URL http://psycnet.apa.org/journals/edu/52/3/117/.

Andrea L Patalano and Colleen M Seifert. Memory for impasses during problem solving. *Memory & Cognition*, 22(2):234–242, 1994.

Paul B Paulus and Mary T Dzindolet. Social influence processes in group brainstorming. *Journal of Personality and Social Psychology*, 64(4):575, 1993.

Paul B Paulus, Nicholas W Kohn, and Lauren E Arditti. Effects of quantity and quality instructions on brainstorming. *The Journal of Creative Behavior*, 45(1):38–46, 2011.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *EMNLP*, volume 14, pages 1532–1543, 2014.

A Terry Purcell and John S Gero. Design and other types of fixation. *Design Studies*, 17(4): 363–383, 1996.

Katharina Reinecke and Krzysztof Z. Gajos. Labinthewild: Conducting large-scale online experiments with uncompensated samples. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*, CSCW '15, pages 1364–1378, New York, NY, USA, 2015. ACM. ISBN 978-1-4503-2922-4. doi: 10.1145/2675133.2675246.

Christoph Riedl, Ivo Blohm, Jan Marco Leimeister, and Helmut Krcmar. Rating scales for collective intelligence in innovation communities: Why quick and easy decision making does not get it right. 2010.

E. F. Rietzschel, B. A. Nijstad, and W. Stroebe. Relative accessibility of domain knowledge and creativity: The effects of knowledge activation on the quantity and originality of generated ideas. *Journal of Experimental Social Psychology*, 43(6):933–946, 11 2007a. ISSN 00221031.

Eric F. Rietzschel, Bernard A. Nijstad, and Wolfgang Stroebe. Relative accessibility of domain knowledge and creativity: The effects of knowledge activation on the quantity and originality of generated ideas. *Journal of Experimental Social Psychology*, 43(6):933–946, 2007b. ISSN 00221031. doi: 10.1016/j.jesp.2006.10.014.

Daniel Ritchie, Ankita Arvind Kejriwal, and Scott R Klemmer. d. tour: Style-based exploration of design example galleries. In *Proc. UISTâĂŹ11*, pages 165–174. ACM, 2011.

Daniel M Russell, Mark J Stefik, Peter Pirolli, and Stuart K Card. The cost structure of sensemaking. In *Proceedings of the INTERACT'93 and CHI'93 conference on Human factors in computing systems*, pages 269–276. ACM, 1993.

Daniel M Russell, Malcolm Slaney, Yan Qu, and Mave Houston. Being literate with large document collections: Observational studies and cost structure tradeoffs. In *System Sciences, 2006. HICSS'06. Proceedings of the 39th Annual Hawaii International Conference on*, volume 3, pages 55–55. IEEE, 2006.

Matthew J Salganik and Karen EC Levy. Wiki surveys: Open and quantifiable social data collection. *arXiv preprint arXiv:1202.0500*, 2012.

R Keith Sawyer. *Explaining creativity: The science of human innovation*. Oxford University Press, 2011.

Thimo Schulze, Marta Indulska, David Geiger, and Axel Korthaus. Idea assessment in open innovation: A state of practice. *ECIS 2012 Proceedings*, pages Paper–149, 2012.

Daniel L Schwartz and Taylor Martin. Inventing to prepare for future learning: The hidden efficiency of encouraging original student production in statistics instruction. *Cognition and Instruction*, 22(2):129–184, 2004.

Daniel L Schwartz, Catherine C Chase, Marily A Oppezzo, and Doris B Chin. Practicing versus inventing with contrasting cases: The effects of telling first on learning and transfer. *Journal of Educational Psychology*, 103(4):759, 2011.

C. M. Seifert, D. E. Meyer, N. Davidson, A. L. Patalano, and I. Yaniv. Demystification of cognitive insight- opportunistic assimilation and the prepared-mind perspective. In R. J. Sternberg and J. E. Davidson, editors, *The Nature of Insight*, volume 124, pages 65–124. MIT Press, Cambridge, MA, 1995.

Juliet P. Shaffer. Multiple hypothesis-testing. *Annual Review of Psychology*, 46:561–584, 1995.

Moushumi Sharmin and Brian P Bailey. Reflectionspace: an interactive visualization tool for supporting reflection-on-action in design. In *Proceedings of the 9th ACM Conference on Creativity & Cognition*, pages 83–92. ACM, 2013.

Morgan M. Shepherd, Robert O. Briggs, Bruce A. Reinig, Jerome Yen, and Jay F. Nunamaker Jr. Invoking social comparison to improve electronic brainstorming: Beyond anonymity. *Journal of Management Information Systems*, 12(3):155–170, 1995.

Pao Siangliulue, Kenneth C. Arnold, Krzysztof Z. Gajos, and Steven P. Dow. 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*, CSCW '15, pages 937–945, New York, NY, USA, 2015a. ACM. ISBN 978-1-4503-2922-4. doi: 10.1145/2675133.2675239. URL http://doi.acm.org/10.1145/2675133.2675239.

Pao Siangliulue, Joel Chan, Krzysztof Z. Gajos, and Steven P. Dow. Providing timely examples improves the quantity and quality of generated ideas. In *Proceedings of the 2015 ACM SIGCHI Conference on Creativity and Cognition*, C&C '15, pages 83–92, New York, NY, USA, 2015b. ACM. ISBN 978-1-4503-3598-0. doi: 10.1145/2757226.2757230. URL http://doi.acm.org/10.1145/2757226.2757230.

Pao Siangliulue, Joel Chan, Steven P Dow, and Krzysztof Z Gajos. Ideahound: Improving large-scale collaborative ideation with crowd-powered real-time semantic modeling. In *Proceedings of the 29th Annual Symposium on User Interface Software and Technology*, pages 609–624. ACM, 2016.

Jasjit Singh and Lee Fleming. Lone inventors as sources of breakthroughs: Myth or reality? *Management Science*, 56(1):41–56, 2010.

Ut Na Sio, Kenneth Kotovsky, and Jonathan Cagan. Fixation or inspiration? A meta-analytic review of the role of examples on design processes. *Design Studies*, 39:70–99, July 2015. ISSN 0142-694X. doi: 10.1016/j.destud.2015.04.004. URL http://www.sciencedirect.com/science/article/pii/S0142694X15000290.

Steven M Smith, Thomas B Ward, and Jay S Schumacher. Constraining effects of examples in a creative generation task. *Memory & Cognition*, 21(6):837–845, 1993.

Jerry O Talton, Daniel Gibson, Lingfeng Yang, Pat Hanrahan, and Vladlen Koltun. Exploratory modeling with collaborative design spaces. *ACM Transactions on Graphics-TOG*, 28(5):167, 2009.

Omer Tamuz, Ce Liu, Serge Belongie, Ohad Shamir, and Adam Tauman Kalai. Adaptively Learning the Crowd Kernel. *arXiv.org*, May 2011.

Y Tanaka, Y Sakamoto, and Takashi Kusumi. conceptual combination versus critical combination: devising creative solutions using the sequential application of crowds. In *Proceedings of the 33rd Annual Conference of the Cognitive Science Society*, 2011.

Loren Terveen, Will Hill, and Brian Amento. Constructing, organizing, and visualizing collections of topically related web resources. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 6(1):67–94, 1999.

Ellis Paul Torrance. *Torrance tests of creative thinking*. Personnel Press, Incorporated, 1968.

Ian Tseng, Jarrod Moss, Jonathan Cagan, and Kenneth Kotovsky. The role of timing and analogical similarity in the stimulation of idea generation in design. *Design Studies*, 29(3): 203–221, 2008.

Barbara Tversky, Maneesh Agrawala, Julie Heiser, Paul Lee, Pat Hanrahan, Doantam Phan, Chris Stolte, and Marie-Paule Daniel. Cognitive design principles for automated generation of visualizations. *Applied spatial cognition: from research to cognitive technology*, pages 53–75, 2006.

L van der Maaten and K Weinberger. Stochastic triplet embedding. *Machine Learning for Signal Processing (MLSP), 2012 IEEE International Workshop on*, pages 1–6, 2012.

Laurens Van Der Maaten. Accelerating t-sne using tree-based algorithms. *The Journal of Machine Learning Research*, 15(1):3221–3245, 2014.

Oana Vuculescu and Carsten Bergenholtz. How to Solve Problems with Crowds: A Computer-Based Simulation Model. *Creativity and Innovation Management*, 23(2):121–136, June 2014. ISSN 1467-8691. doi: 10.1111/caim.12059. URL http://onlinelibrary.wiley.com/doi/10.1111/caim.12059/abstract.

Thomas B Ward. Structured imagination: The role of category structure in exemplar generation. *Cognitive Psychology*, 27(1):1–40, 1994.

Bernhard Weber and Guido Hertel. Motivation gains of inferior group members: a meta-analytical review. *Journal of personality and social psychology*, 93(6):973, 2007.

Sarah Weir, Juho Kim, Krzysztof Z Gajos, and Robert C Miller. Learnersourcing subgoal labels for how-to videos. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*, pages 405–416. ACM, 2015.

Thomas N. Wisdom and Robert L. Goldstone. Innovation, Imitation, and Problem Solving in a Networked Group. *Nonlinear Dynamics-Psychology and Life Sciences*, 15(2):229, 2011.

Anbang Xu and Brian Bailey. A reference-based scoring model for increasing the findability of promising ideas in innovation pipelines. In *Proc. CSCWâĂŹ12*, pages 1183–1186. ACM, 2012.

Lixiu Yu and Jeffrey Nickerson. Cooks or cobblers?: crowd creativity through combination. *Proc. CHIâĂŹ11*, 2011.

Lixiu Yu, Aniket Kittur, and Robert E. Kraut. Searching for analogical ideas with crowds. In *Proceedings of the 32Nd Annual ACM Conference on Human Factors in Computing Systems*, CHI '14, pages 1225–1234, New York, NY, USA, 2014a. ACM. ISBN 978-1-4503-2473-1. doi: 10.1145/2556288.2557378. URL http://doi.acm.org/10.1145/2556288.2557378.

Lixiu Yu, Aniket Kittur, and Robert E Kraut. Distributed analogical idea generation: inventing with crowds. In *Proceedings of the 32nd annual ACM conference on Human factors in computing systems*, pages 1245–1254. ACM, 2014b.

Yisong Yue, Chong Wang, Khalid El-Arini, and Carlos Guestrin. Personalized collaborative clustering. In *Proceedings of the 23rd international conference on World wide web*, pages 75–84. ACM, 2014.

Liang Zeng, Robert W. Proctor, and Gavriel Salvendy. Fostering creativity in product and service development: validation in the domain of information technology. *Hum Factors*, 53 (3):245–70, 6 2011. ISSN 0018-7208.

Appendix A

Birthday Message Ideation Task Instruction

The figure below shows the instructions we used in the ideation task in Chapter 4.

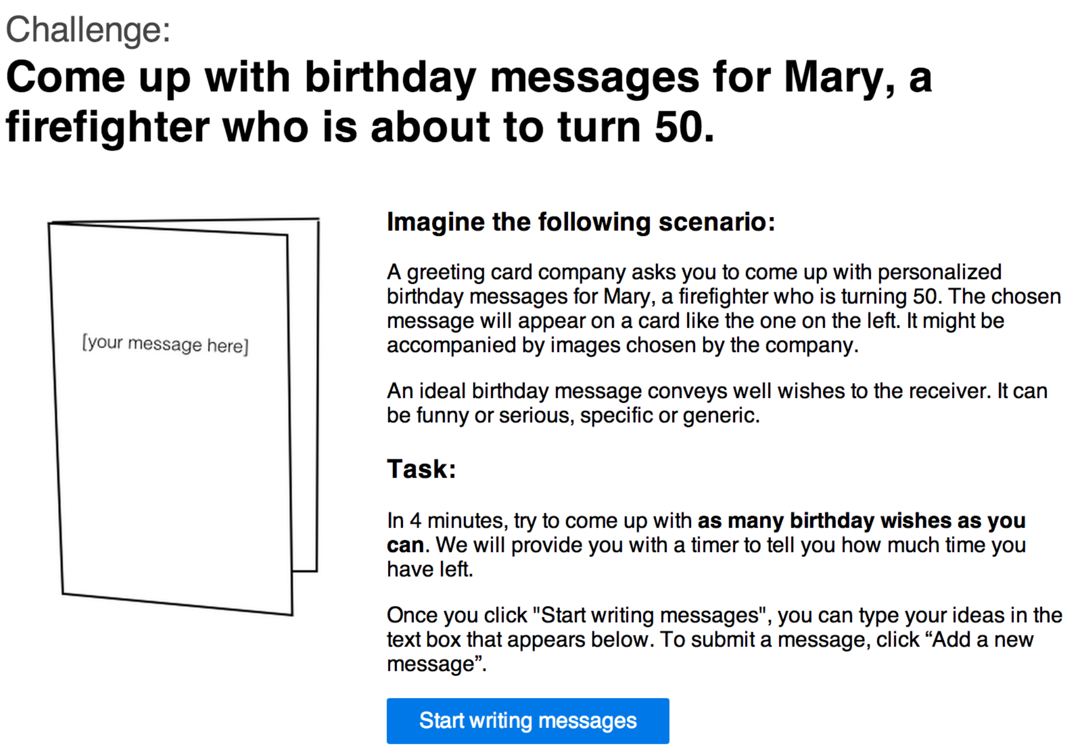


Figure A.1: Instruction for the task used in the experiment

Appendix B

Seed Ideas for Solution Synthesis

The list below shows ideas that were used as seed ideas and their corresponded groups for the experiment in Chapter 6. Participants did not see the labels of the groups. For brevity, the list only shows the title of the ideas. You can find full description of each idea on the OpenIDEO website[1].

1. Recruit through curiosity about self
   (a) The Grand Family Tree
2. Recruit through in-person social events
   (a) Swab parties (on the basis of Tupperware parties)
   (b) Tributes & Toasts
3. Piggyback from medical-related activities
   (a) Check it out at your check up
   (b) If you're at the Dentist with your mouth open anyways
4. Make use of waiting time
   (a) Department of Motor Vehicles AND Marrow
   (b) I'm bored! Time to sign up
5. Recruit people through their companies
   (a) Corporate Swabbing
   (b) Get Companies to Push Employee Swabbing as CSR
   (c) Sign up bone marrow registry, get extra vacation days!
6. Publicize in public space

[1] https://challenges.openideo.com/challenge/how-might-we-increase-the-number-of-bone-marrow-donors-to-help-save-more-lives/concepting

   (a) Educate while you...well, educate in the restroom.
   (b) Flush for Life
   (c) Art and Awareness
   (d) MythBUSters! - Destroying the Myths in a BIG way
7. Recruit prisoners
   (a) Bones for freedom.
   (b) can we work with ex-prisoners?
   (c) Prisoner Bone Marrow Registry
   (d) Save a Life, Earn a Life
8. International registration
   (a) Beyond the familiar: Spreading the word Globally
   (b) Bone Marrow Without Borders
   (c) Go beyond the familiar: make the Campaign really Global
   (d) Simplify international donations
9. Ask people who are already giving
   (a) Ask and You Shall Receive . . . If You Ask the RIGHT Person
   (b) Blood Donors Dummy- Teach them that Marrow matters too!
   (c) Military to the bone
   (d) Turning Dollar Donors into Marrow Donors
   (e) Uncle Sam Wants Your Marrow
   (f) We can be heroes
10. Ask for help from celebrities
   (a) 15 year old boy + famous singer + story telling = spreading the word
   (b) Celebrity Goes Through Donation Process
   (c) Cricket Star ads to encourage South Asian donors
   (d) Enchantment time with Guy Kawasaki
   (e) Fashionable Donors
   (f) Got a call from Shaq and said "YES"
11. Launch publicizing campaigns
   (a) Nick naming Bone Marrow
   (b) This Isn't Your Mama's Bone Marrow Donation Process
   (c) Bucket List

    (d) Campaign: ‘Super hero’s genes are in your marrow’
    (e) Colored bone relays tracked on a website - endless/free campaign
    (f) The bone marrow tour!
    (g) “It Gets Marrow” - Stories of Bone Marrow Donnors and Recipients

12. Recruit through religion institutes
    (a) Assuaging African-American Fears Through Clergy-Based Education
    (b) Getting help from the religious community
    (c) Give a lifetime for Eid/Diwali!
    (d) Places of Worship
    (e) Ready Sources
    (f) Take a benefit from those times and places bring sympathy to society
    (g) Tap into the Influence of Spiritual Leaders

13. Recruit through universities
    (a) Bone Marrow donation signup counters in University and College Health centers
    (b) Engaging undergraduate students in campaigning for bone marrow donation
    (c) match up | one week . 100k cheeks
    (d) University-Based Open Innovation Awareness Websites built on a DIY Kit
    (e) Utilizing the power of student clubs
    (f) Train the Trainers. Go Team!
    (g) Bone Marrow Advo Kit: Comprehensive Training Curriculum

14. Inform and recruit through movies or TV shows
    (a) Bone Marrow Donation as Reality TV
    (b) Dr. Oz appeal
    (c) Hollywood for Bone Marrow Donation
    (d) Lisa Simpson promotes bone marrow donation
    (e) Pop Culture Creates Pop Action
    (f) Reality Bites
    (g) TV game
    (h) The Donation - A Feature Film On Donors and Recipients
    (i) Sweeping the Clouds Away

15. Make use of digital social network
    (a) 6 Degrees of Bone Marrow Seperation

(b) a handwritten letter posted on facebook
(c) "Become a Life-Saver" - Social Bone Marrow Network
(d) Connect In: An Online Network to Help Cancer Patients
(e) I "Like" Bone Marrow Donation, Do you?
(f) Instant Karma
(g) Just click 'Join'. Facebook sign u
(h) Social Medical Network
(i) Meet a Recipient. Save a Life. Tell your story.
(j) Celebrate Hero Anniversaries on Social Media
(k) Virtual Bone Marrow Registration

16. Teaming with companies
    (a) Make it a Game
    (b) The Q-tip Prize
    (c) GOOGLIZE IT
    (d) B isn't just for Becks/beer
    (e) Know Your Type
    (f) Jones Soda and M&Ms
    (g) Q-Tip Partnership
    (h) Save a Life and Save Some Gas Money
    (i) Cells for stem cells
    (j) Credits from health insurance for BM donation
    (k) Donor Discounts!
    (l) Discount Genotyping with Bone Marrow Registration
    (m) Pharamacy Donation
    (n) Save a Life & Kill a Bill