View metadata, citation and similar papers at core.ac.uk
brought to you by CORE
provided by CiteSeerX

Toward Collaborative Ideation at Scale — Leveraging Ideas from Others to Generate More Creative and Diverse Ideas

Pao Siangliulue[1] **Kenneth C. Arnold**[1] **Krzysztof Z. Gajos**[1] **Steven P. Dow**[2]

[1]Harvard School of Engineering and Applied Sciences
Cambridge, MA USA
{paopow, kcarnold, kgajos}@seas.harvard.edu

[2]Carnegie Mellon University
Pittsburgh, PA USA
spdow@cs.cmu.edu

ABSTRACT

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. We contribute 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.

Author Keywords

Collaborative ideation, inspirational examples, creativity

ACM Classification Keywords

H.5.m. Information Interfaces and Presentation (e.g. HCI): Miscellaneous

INTRODUCTION

The "lone inventor" is a myth: even geniuses benefit from exposure to ideas of others [29]. 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 [14, 19, 5]. By generating ideas together, people can produce more diverse ideas than if each person ideated alone, and this diversity can lead to more creative overall solutions [23, 27].

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 [19, 17] or extract schemas from examples and search for the schema that allows analogical transfer for a new idea [38, 36]. 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. Ideators 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 [15, 30, 22, 16]. A set of uninspiring examples may fixate ideators 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 [22, 26, 20]. Similarly, if the set of examples is diverse (i.e., if the ideas within the set were substantially different from each other),

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from Permissions@acm.org.
CSCW '15, March 14 - 18 2015, Vancouver, BC, Canada
Copyright is held by the owner/author(s). Publication rights licensed to ACM.
ACM 978-1-4503-2922-4/15/03...$15.00
http://dx.doi.org/10.1145/2675133.2675239

the diversity of generated ideas should also increase [25]. 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 [28, 32, 35, 40, 37, 39]. 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 [31, 34] 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 ideators with sets of ideation examples 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.

In this work, we made the following contributions:

- We developed and validated a crowd-powered method for automatically constructing an "idea map" that can be used to extract diverse sets of examples at scale.

- We conducted 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.

PRIOR WORK

How creative and diverse examples affect ideation

Seeing others' ideas can have positive effects on one's own ideation as it can help people come up with ideas that they would not have thought of otherwise. 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 [8, 9, 4]. Exposure to others' ideas also accelerates the generation of ideas across different semantic categories, increasing productivity overall [25].

However, seeing other people's ideas may have unintended side effects. Specifically, people tend to generate ideas that borrow concepts from presented examples [15, 30, 22, 16]. 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.*

Conversely, presenting people with sets of diverse and creative ideas may stimulate pursuit of new directions of thought. Empirical work suggests that exposing people to novel ideas, as opposed to common ones, can result in more novel ideas [22]. 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. Building on this work, we looked into the effects of creativity — which combines novelty and value of ideas — and diversity of examples on idea generation. Our results support this finding: when presented with creative examples, people not only integrate examples into their ideas but also add their own spin to them and innovate in other ways.

The value of exposing people to diverse sets of examples is also supported by cognitive models of creativity. 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 [25, 24]. Ideas generated by others can act as external stimuli, activating material retrieval from memory for idea production. If the examples are diverse, one will retrieve a diverse set of raw material for idea synthesis and will be more likely to generate a diverse set of 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 [7] 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 [25].

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 [26, 20]. One can infer the overall performance of others from ideas that one sees and try to match with ideas of the same calibre. We expect that exposing people to high quality, creative examples generated by peers will raise their aspirations while showing them less creative examples would likely lower the quality of subsequent ideas.

To summarize, prior research suggests that exposing people to ideas of others may positively impact one's own ideation outcomes especially if 1) each idea is individually of high quality, meaning that it is both novel and appropriate; 2) the *set* of ideas that a person is exposed to covers many semantic categories.

These insights lead us to propose 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, we hypothesize that both of these interventions will increase both the *creativity* and the *diversity* of ideas generated. Our research seeks to integrate appropriate methods for applying these interventions at scale in a way that does not require expert intervention.

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 [28]. In other works where more refined measures are needed, users rate creativity of ideas on different Likert scales [40, 32]. Xu and Bailey demonstrated a mechanism that helps ensure that voting results from non-experts match those of experts [35].

Scalably assessing the diversity of a set of ideas, however, has not been as well studied. We next dive into prior work on this branch of research.

Prior approaches for assessing diversity

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

Semantic categories

Manually created semantic categories have been used in prior research in creativity, either as ways to select a diverse set of ideas [25], or as a way to evaluate the diversity of creative artifacts [10, 15].

Efficient crowd-based mechanisms exist to label large collections of items with semantic categories or tags. Some take

Figure 1. An idea map generated by our system, showing emergent clusters of ideas around different themes and sentiments.

the approach of generating labels or tags for each individual item [18], while others produce hierarchical taxonomies capturing the semantic structure of the concepts represented in the item set [3, 2, 1]. Differences in contributors' mental models have been a persistent difficulty in semantic categorization even for experts [3]. A complete system for organizing ideas should include elements of both discrete semantic categories and continuous quantitative similarity, but because of difficulties we encountered with categorization approaches in pilot experiments, we chose to focus on continuous similarity in this work. Compared to these methods, our approach offers more fine-grained assessment of similar ideas.

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 [6]. 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, like ours, 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 [13, 11] or a Euclidean space [31, 34]. 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. In the next section, we detail how we build our method on top of these non-semantic techniques.

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 cap-

ture only superficial similarity [1]. 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 [31] that uses triplet similarity comparisons to place ideas in a two-dimensional map. The map is 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 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 [31]. We use t-Distributed Stochastic Triplet Embedding (t-STE) [34] 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 [31] 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.

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 the appendix. 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 [12, 33, 30, 22, 21]. 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.

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.

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 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!"*

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

Validating the Idea Map

To validate the diversity ratings created by the idea map, we collected similarity ratings [6] 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.

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.

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.)

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.

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 [6]. 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.

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.

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.

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 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 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 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 2d) also produced no significant effect (F(2,41) = 1.56, n.s.)

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 [22, 26, 20]), 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

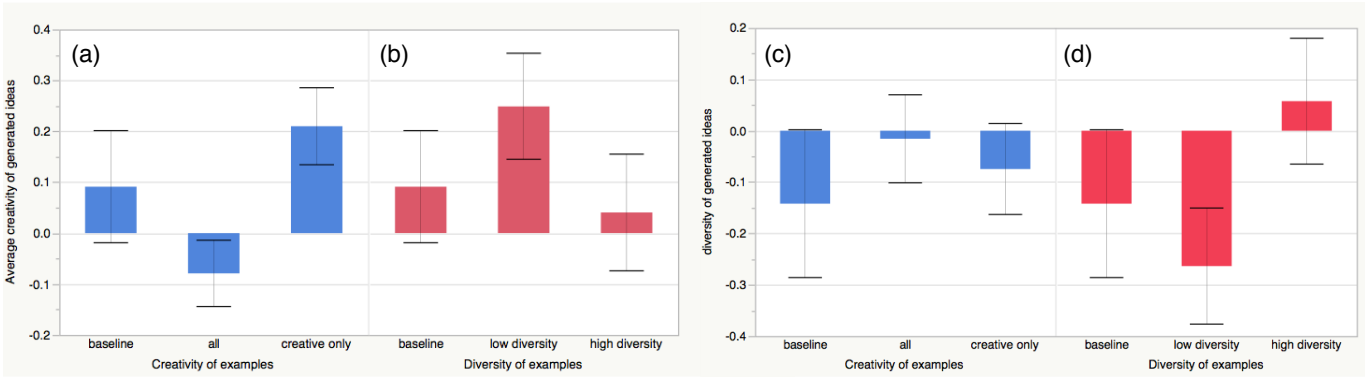

Figure 2. (a) Participants in the Creative Only condition generated more creative ideas than participants in the All condition. (b) There is no difference in average creativity of generated ideas across groups seeing different levels of diversity. (c) There is no difference in the diversity of generated 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.

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* [24] 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 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 3b). In both interventions, the similarity to examples was lower than the self-fixation observed in the baseline condition (M=0.79).

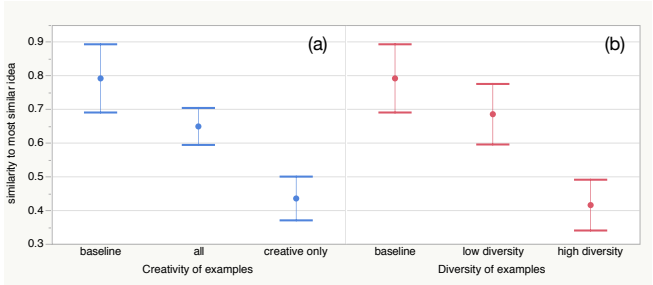

Figure 3. (a) Participants in the Creative only condition were less fixated than those in the baseline and the All examples condition. (b) Participants who saw a set of diverse examples were less fixated those in the baseline condition and those who saw a set of examples with low diversity.

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.

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 [26, 20]. While the two explanations involve very different mechanisms, they both support the value of presenting ideators 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 [37]), 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.

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 work 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 [28, 40, 32, 35], 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 ideators.

ACKNOWLEDGMENTS

This work was funded in part by a Sloan Research Fellowship, gifts from Google and Adobe and awards from the National Science Foundation (IIS-1208382 and IIS-1217096). We thank Ofra Amir and Joel Chan for feedback on the manuscript.

APPENDIX

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

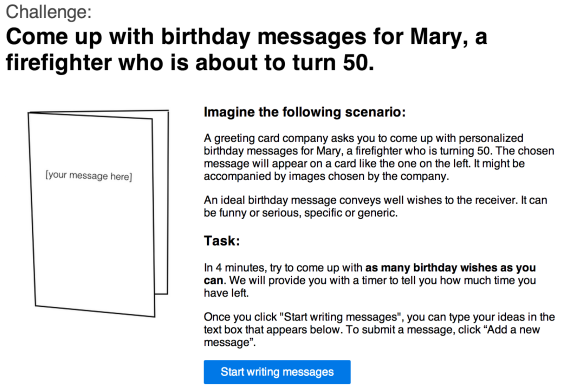

Figure 4. Instruction for the task used in the experiment

REFERENCES

1. André, P., Kittur, A., and Dow, S. P. 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, ACM (New York, NY, USA, 2014), 989–998.
2. Chilton, L. B., Kim, J., André, P., Cordeiro, F., Landay, J. A., Weld, D. S., Dow, S. P., Miller, R. C., and Zhang, H. Frenzy: Collaborative data organization for creating conference sessions. In *Proceedings of the 32Nd Annual ACM Conference on Human Factors in Computing Systems*, CHI '14, ACM (New York, NY, USA, 2014), 1255–1264.
3. Chilton, L. B., Little, G., Edge, D., Weld, D. S., and Landay, J. A. Cascade: Crowdsourcing taxonomy creation. In *Proceedings of the 2013 ACM annual conference on Human factors in computing systems*, ACM (2013), 1999–2008.
4. Dennis, A. R., and Valacich, J. S. Computer brainstorms: More heads are better than one. *Journal of Applied Psychology 78*, 4 (1993), 531.
5. Dow, S., Fortuna, J., Schwartz, D., Altringer, B., Schwartz, D., and Klemmer, S. 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, ACM (New York, NY, USA, 2011), 2807–2816.
6. Dow, S., Glassco, A., Kass, J., Schwarz, M., Schwartz, D., and Klemmer, S. Parallel prototyping leads to better design results, more divergence, and increased self-efficacy. *Transactions on Computer-Human Interaction (TOCHI 17*, 4 (2010).
7. Finke, R. A., Ward, T. B., and Smith, S. M. *Creative cognition: Theory, research, and applications*. MIT press Cambridge, MA, 1992.
8. Gallupe, R. B., Bastianutti, L. M., and Cooper, W. H. Unblocking brainstorms. *Journal of Applied Psychology 76*, 1 (1991), 137.
9. Gallupe, R. B., Dennis, A. R., Cooper, W. H., Valacich, J. S., Bastianutti, L. M., and Nunamaker, J. F. Electronic brainstorming and group size. *Academy of Management Journal 35*, 2 (1992), 350–369.
10. Goldenberg, O., Larson, J. R., and Wiley, J. Goal instructions, response format, and idea generation in groups. *Small Group Research 44*, 3 (2013), 227–256.
11. Gomes, R. G., Welinder, P., Krause, A., and Perona, P. Crowdclustering. In *Advances in Neural Information Processing Systems 24*, J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira, and K. Weinberger, Eds. Curran Associates, Inc., 2011, 558–566.
12. Guilford, J. P. The nature of human intelligence.
13. Heikinheimo, H., and Ukkonen, A. The Crowd-Median Algorithm. *First AAAI Conference on Human ...* (2013), 69–77.
14. Herring, S., Chang, C.-C., Krantzler, J., and Bailey, B. Getting inspired!: understanding how and why examples are used in creative design practice. *CHI '09: Proceedings of the 27th international conference on Human factors in computing systems* (2009).
15. Jansson, D. G., and Smith, S. M. Design fixation. *Design Studies 12*, 1 (1991), 3–11.
16. Kohn, N. W., and Smith, S. M. Collaborative fixation: Effects of others' ideas on brainstorming. *Applied Cognitive Psychology 25*, 3 (2011), 359–371.
17. Kumar, R., Satyanarayan, A., Torres, C., Lim, M., Ahmad, S., Klemmer, S. R., and Talton, J. O. Webzeitgeist: design mining the web. In *CHI '13: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ACM Request Permissions (Apr. 2013).
18. Law, E., and Von Ahn, L. Input-agreement: a new mechanism for collecting data using human computation games. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ACM (2009), 1197–1206.
19. Lee, B., Srivastava, S., Kumar, R., Brafman, R., and Klemmer, S. R. Designing with interactive example galleries. In *CHI '10: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ACM Request Permissions (Apr. 2010).
20. Leggett Dugosh, K., and Paulus, P. B. Cognitive and social comparison processes in brainstorming. *Journal of Experimental Social Psychology 41*, 3 (May 2005), 313–320.
21. Lewis, S., Dontcheva, M., and Gerber, E. Affective computational priming and creativity. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ACM (2011), 735–744.
22. Marsh, R. L., Landau, J. D., and Hicks, J. L. How examples may (and may not) constrain creativity. *Memory & Cognition 24*, 5 (1996), 669–680.
23. Nagasundaram, M., and Dennis, A. R. When a group is not a group the cognitive foundation of group idea generation. *Small Group Research 24*, 4 (1993), 463–489.
24. Nijstad, B. A., and Stroebe, W. How the group affects the mind: A cognitive model of idea generation in groups. *Personality and Social Psychology ...* (2006).
25. Nijstad, B. A., Stroebe, W., and Lodewijkx, H. F. M. Cognitive stimulation and interference in groups: Exposure effects in an idea generation task. *Journal of Experimental Social Psychology 38*, 6 (2002), 535–544.
26. Paulus, P. B., and Dzindolet, M. T. Social influence processes in group brainstorming. *Journal of Personality and Social Psychology 64*, 4 (1993), 575.
27. Paulus, P. B., Kohn, N. W., and Arditti, L. E. Effects of quantity and quality instructions on brainstorming. *The Journal of Creative Behavior 45*, 1 (2011), 38–46.
28. Salganik, M. J., and Levy, K. E. Wiki surveys: Open and quantifiable social data collection. *arXiv preprint arXiv:1202.0500* (2012).
29. Singh, J., and Fleming, L. Lone inventors as sources of breakthroughs: Myth or reality? *Management Science 56*, 1 (2010), 41–56.
30. Smith, S. M., Ward, T. B., and Schumacher, J. S. Constraining effects of examples in a creative generation task. *Memory & Cognition 21*, 6 (1993), 837–845.
31. Tamuz, O., Liu, C., Belongie, S., Shamir, O., and Kalai, A. T. Adaptively Learning the Crowd Kernel. *arXiv.org* (May 2011).
32. Tanaka, Y., Sakamoto, Y., and Kusumi, T. 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).
33. Torrance, E. P. *Torrance tests of creative thinking*. Personnel Press, Incorporated, 1968.
34. van der Maaten, L., and Weinberger, K. Stochastic triplet embedding. *Machine Learning for Signal Processing (MLSP), 2012 IEEE International Workshop on* (2012), 1–6.
35. Xu, A., and Bailey, B. A reference-based scoring model for increasing the findability of promising ideas in innovation pipelines. In *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work*, ACM (2012), 1183–1186.
36. Yu, L., Kittur, A., and Kraut, R. E. Distributed analogical idea generation: inventing with crowds. In *Proceedings of the 32nd annual ACM conference on Human factors in computing systems*, ACM (2014), 1245–1254.
37. Yu, L., Kittur, A., and Kraut, R. E. Distributed analogical idea generation: Inventing with crowds. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '14, ACM (New York, NY, USA, 2014), 1245–1254.
38. Yu, L., Kittur, A., and Kraut, R. E. Searching for analogical ideas with crowds. In *Proceedings of the 32Nd Annual ACM Conference on Human Factors in Computing Systems*, CHI '14, ACM (New York, NY, USA, 2014), 1225–1234.
39. Yu, L., Kittur, A., and Kraut, R. E. Searching for analogical ideas with crowds. In *Proceedings of the 32Nd Annual ACM Conference on Human Factors in Computing Systems*, CHI '14, ACM (New York, NY, USA, 2014), 1225–1234.
40. Yu, L., and Nickerson, J. Cooks or cobblers?: crowd creativity through combination. *CHI '11: Proceedings of the 2011 annual conference on Human factors in computing systems* (2011).