CrowdMuse: Supporting Crowd Idea Generation through User Modeling and Adaptation

Victor Girotto
School of Computing, Informatics, and Decision Systems Engineering
Arizona State University
Tempe, AZ, USA
victor.girotto@asu.edu

Erin Walker
School of Computing and Information; LRDC
University of Pittsburgh,
Pittsburgh, PA, USA
eawalker@pitt.edu

Winslow Burleson
Rory Meyers College of Nursing
New York University
New York, NY, USA
wb50@nyu.edu

ABSTRACT
Online crowds, with their large numbers and diversity, show great potential for creativity. Research has explored different ways of augmenting their creative performance, particularly during large-scale brainstorming sessions. Traditionally, this comes in the form of showing ideators some form of inspiration to get them to explore more categories or generate more and better ideas. The mechanisms used to select which inspirations are shown to ideators thus far have not taken into consideration ideators' individualities, which could hinder the effectiveness of support. In this paper, we introduce and evaluate CrowdMuse, a novel adaptive system for supporting large-scale brainstorming. The system models ideators based on their past ideas and adapts the system views and inspiration mechanism accordingly. We evaluate CrowdMuse over two iterative large online studies and discuss the implication of our findings for designing adaptive creativity support systems.

Author Keywords
Creativity; brainstorming; crowd; adaptive systems.

ACM Classification Keywords
• Information systems→ Collaborative and social computing systems and tools • Human-centered computing→ Collaborative and social computing systems and tools

INTRODUCTION
Online crowds show great potential for creativity. This is in great part due to the large numbers and diversity of participants [8,11]. In small groups, the main contributor to an increased performance is synergy, that is, when one person builds on ideas proposed by others [8]. These synergistic ideas would hardly occur in individual ideation. Therefore, one might expect that by adding hundreds more people to ideation, the likelihood of synergy happening would only increase.

Nonetheless, simply recruiting large numbers of ideators is not enough to ensure a creative output. The same scale and diversity that can boost ideation also presents challenges that hinder the creative output of crowds. The sheer amount of ideas generated can hinder synergistic performance, since an individual is unlikely to be able to read all of the ideas (thus possibly missing the one that could inspire them), much less pay attention to them, which is a requirement for influence [10,18]. Therefore, large-scale brainstorming sessions need to be appropriately designed and supported.

Research has attempted to do that in different ways, usually by withholding the entire solution space (all the ideas generated so far) and only exposing ideators to *inspirations*—usually a short text snippet meant to inspire further ideas. These inspirations have taken different forms. For example, Chan, Dang, & Dow [4] employed facilitators to generate inspirations (e.g. questions to promote reflection) during an ideation session. Siangliulue et al. [23] attempted to inspire ideators by showing them a small set of ideas chosen either for their diversity (ideas that differ significantly among themselves) or creativity. Finally, in our previous work we attempted to increase the effect of inspirations by adding a small task (e.g. rating the idea) to boost attention to the ideas [12].

The common thread between these examples is that they focus on the kind of inspiration being shown rather than on the ideator it is being shown to. In other words: should the same inspiration be presented to two different ideators? Could individual differences between ideators affect how effective an inspiration is? The creativity literature points towards an answer to these questions. Theoretical models of idea generation propose that individuals differ on which concepts or categories they generate ideas [3,18]. This means that each ideator is more likely to focus on some areas (i.e. idea categories) than others. An inspiration strategy that does not take these individualities into consideration may be missing out on leveraging ideators' unique strengths. For example, if an ideator is more familiar with ideas in category

C&C '19, June 23–26, 2019, San Diego, CA, USA

ACM 978-1-4503-5917-7/19/06…$15.00
https://doi.org/10.1145/3325480.3325497

A than those in B, showing ideas in category B may not effectively inspire him or her to come up with new ideas.

In this paper, we explore how to tailor inspiration selection to individual ideators. The overarching research question explored here is: "*How can we adapt inspirations to ideators in order to improve ideation performance?*" To do so, we present CrowdMuse, a system that models ideators based on their categories of ideation and adapts itself to improve creative performance. We begin by describing the literature that inspired this work. We then introduce CrowdMuse, describing how it functions in detail, and relating its design back to the literature. We finish by describing two large-scale online studies in which we evaluate CrowdMuse and its adaptive mechanism. We make the following contributions:

- We introduce CrowdMuse and the methods it uses to model ideators and adapt to them;
- We validate the system's effectiveness method in two studies, demonstrating that an adaptive system can improve the breadth of ideation, given an appropriate categorization of the inspiration pool.

RELATED WORK

Idea Generation Models

There are two models of idea generation central to this paper. The first model is the Search for Ideas in Associative Memory (SIAM) [17–19]. This model describes idea generation in terms of memory recall. The assumption is that there are two memory systems: working memory (WM) and long-term memory (LTM). LTM is essentially unlimited, and is organized in images, which are central concepts (e.g. a computer) along with associated features (e.g. has a CPU, has a storage unit). WM is where conscious processing takes place but is quite limited. An additional component worth noting is the search cue, existing in the WM, which serves as the cue to search the LTM. This search cue is comprised of items such as the problem definition, previous ideas, or personal experiences.

With these components in place, idea generation is then described in terms of two loops between the two memory systems. The first loop is the image retrieval loop, where an image is retrieved from LTM and loaded into WM. At this point, the image and associated features are available as the basis for the second loop, the idea production loop. The individual can now produce ideas using the image, its features, and whatever is on the search cue. This goes on until no more ideas can be thought of using the current image. At this point, the individual reverts back to the first loop, searching for another image to be loaded into WM. This process follows until no more images can be retrieved, ending idea generation.

The second theoretical model is the matrix model [3,21]. Operating at the level of categories of ideas, it follows the notion that there are categories that are more or less likely to be visited by an ideator. This concept is represented through a matrix of category transition probabilities. Rows and columns represent different categories. Each cell contains a number between 0 and 1, representing the probability of transitioning from a category (row) to another (column). The diagonal of the matrix, therefore, represents the probability of staying within the same category (similar to SIAM's idea production loop), and the other cells represent the probability of transitioning to a different category (similar to SIAM's image retrieval loop). While this model does not explain the flow underlying idea generation in such detail as SIAM does, its matrix representation makes the transition between categories more concrete. In this work, we make use of both models to inform our design and adaptive mechanisms, as will be described later.

Enhancing Crowd-Scale Brainstorming

Given the potential for promoting the creative potential of crowds, research has already examined ways of supporting brainstorming sessions at a large scale. This has been done through a set of distinct approaches.

At the most basic level, brainstorming has been enhanced simply by showing other ideas to ideators. The approaches differ in how they choose the ideas or how they show them. For example, selecting a set of diverse or creative ideas can improve their effects over random sets of ideas [23]. The timing and delivery method of these inspirations can also affect their efficacy, showing benefits to giving users a choice of when to receive inspirations or in employing a smart strategy for choosing the right moment to do so [25]. Finally, increasing the attention to the inspirations ideas, such as by asking questions about the inspirations, can also improve performance under certain circumstances [12].

Since simple exposure to other ideas can bring its own set of issues (e.g. fixation [15] or limitation of the number of categories surveyed [22]), others have proposed ways of inspiring users through abstractions of other ideas or features of the problem. For example, previous work has found some advantage to using machine-generated abstractions of others' ideas as an inspiration [5]. Alternatively, crowd workers can be used to identify and generate schemas to be used as inspirations [27,28]. Another way in which abstractions can be used is through real-time facilitators. This was tested by Chan, Dang, & Dow [4] through their IdeaGens system. They found that by using stimulating strategies such as simulations (asking ideators to imagine scenarios), facilitators can improve ideator's fluency and creativity. Features of the ideation problem can also contribute, such as by identifying domains of expertise relevant to it, or presenting ideators with constraints [29,30].

A final approach considered is that of highly structured human-powered processes. For example, it has been shown that a human-powered genetic algorithm, in which ideas are mixed and selected through several iterations, can result in greater creativity of later ideas [31]. Even more structured, BlueSky employs a crowd-powered algorithm to evenly contribute to the solution space and reduce duplicates [14].

Figure 1 The CrowdMuse system has two main views: the idea workspace (1) allows users to view and manipulate ideas by hovering over them (2); and the solution space (3) provides an overview of the density of ideas developed for each tag.

Current approaches have not yet considered adapting inspirations to individual ideators according to the models previously discussed. This work aims to take a first step in filling this gap by proposing one way of applying theories of creativity to ideator modelling and adaptation.

THE CROWDMUSE SYSTEM

CrowdMuse, (Figure 1) is comprised of two main views. The first, on the left, is the **idea workspace** (#1). The purpose for this view is to allow users to explore and manipulate existing ideas. At the top of the view, a toolbar displays several choices. On its left, there are two buttons, one for displaying all the user's own ideas, the other for displaying the user's favorite ideas. An idea can be favorited by hovering over it and clicking the favorite button (see #2 in Figure 1). At the right of the toolbar, you find a description of what is currently being shown in the workspace (e.g. "Showing *your favorite* ideas" or "Showing *ideas with tag food*", followed by a count of the number of ideas being displayed and a help button (if clicked, a short description of the view is shown).

The workspace enables two other kinds of actions: combining and refining ideas. Ideas can be combined by dragging one idea onto another. This opens a popup showing both ideas, and a space for typing the combined new idea. An idea can also be refined by hovering over it and clicking on the refine button (Figure 1, #2). A popup will then show up with the idea to be refined, allowing the user to edit its text and submit the updated version. These mechanisms were added in accordance to the principles of brainstorming—in which participants are encouraged to build on one another's ideas [20]—and findings from research, which have demonstrated the importance of combinations and subsequent iteration [6,7,9,16,26].

The second view is the **solution space**, occupying most of the right side of the interface (Figure 1, #3). By using a matrix form of visualization [1], the purpose for the solution space is to provide an overview of which categories have been thoroughly explored and, conversely, those which are yet to be explored. This overview is also important so that ideators are not completely blind to other ideators' performance and can at least try to be more consistent with their tagging of ideas. The solution space is represented as an *n x n* matrix in which the rows and columns correspond to the idea categories developed so far. The color of the cell indicates how many ideas have been developed at the intersection of two categories—the darker the cell, the more ideas have been developed within that intersection. Clicking a cell will open all ideas at that category intersection in the idea workspace.

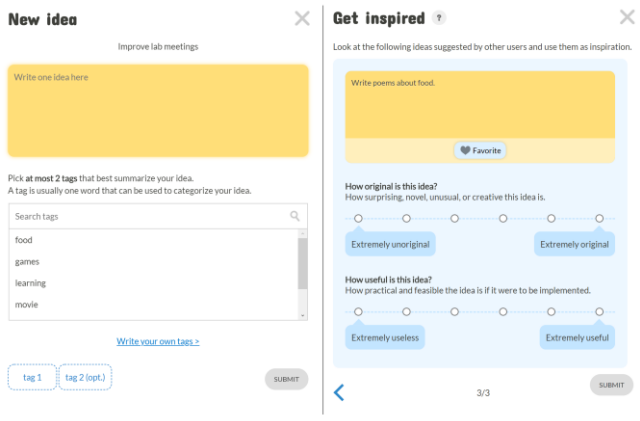

Figure 2 The pop-ups for adding a new idea (left) and when clicking the inspiration button (right).

Users can add ideas by clicking the "new idea" button at the top of the UI. When adding a new idea, the user is prompted to pick at most two categories for the idea (based on previously used categories), or to suggest new ones. To the right of the new idea button is an inspiration button, which when clicked presents three ideas along with a small microtask on each (e.g. "rate the idea's originality and usefulness"). The microtasks are used to increase the attention to ideas and consequently their effect on ideators [12]. This mechanism is similar to other studies described in the literature, employing a pull model where inspirations are requested by ideators through the click of a button [4,12,25]. Figure 2 displays both popups.

Adaptations

The system's purpose is to enhance idea generation by prioritizing categories that could be inspiring to an ideator. Categories in the CrowdMuse system are user-generated groupings of the ideas. This prioritization approach contrasts with current approaches, in which inspiration selection does not take the ideator into consideration, instead being, for example, randomized [12], chronological [4], or focusing on aspects of the inspiration set (e.g. set diversity) [23,24].

As described in the related work, ideators have unique cognitive structures [3,18]. Generally, this means that ideators are more likely to come up with ideas within some categories rather than others, and the ideas may be somewhat temporally clustered together (i.e. ideas of similar categories may be suggested temporarily closer to each other). Therefore, leaving the selection of inspirations to chance may cause them to fail in inspiring (or having as much effect on) ideators due to two factors. First, the ideator is not highly fluent in the chosen categories. This is best visualized with the matrix model. Say, for example, that an ideator has just generated an idea in category A, and from there can switch to categories B, in which she is highly fluent (that is, she has high within-category likelihoods)—and C, in which she is quite inarticulate (that is, she has low within-category likelihoods). In this scenario, an inspiration that touches on category B is much more likely to yield positive results than an inspiration on category C.

The second factor is that the inspiration can break an ideator's train of thought; as proposed by the SIAM model, when ideators generate ideas, they have a concept loaded in their short-term memory (STM). This concept stays loaded until they repeatedly fail to generate more ideas with it. However, if the ideator is exposed to an inspiration that does not match their currently loaded concept, it may interrupt their train of thought, in practice curtailing their fluency within that category (namely, their depth). Therefore, existing research on idea generation shows that inspirations must be carefully chosen to not cause more harm than good.

In practice, the choice of ideas has been shown to influence performance. For example, [23] compared showing random ideas with an explicitly diverse set of ideas, finding the diverse set to yield greater diversity in idea generation. They also found a set of more creative inspirations yielding more creative results. In our previous work, we also found that making the set of inspiration ideas similar among themselves yielded either no or negative effects [12]. The effect dramatically changed when we changed the selection mechanism to be completely randomized, causing the inspirations to improve the breadth of ideation in some cases. Thus, in practice we find that choosing the right ideas for inspirations is important.

Therefore, the CrowdMuse system implements two forms of adaptations: explicit and subtle adaptations. In this subsection, we explain both types of adaptations based on the past literature. How these adaptations are powered (e.g. how does the system choose a new category to suggest) is addressed in the next subsection, User Modeling.

Explicit adaptations are designed to be the most influential form of inspiration. They exist in the inspiration mechanism, which presents users with three ideas, each with an accompanying rating microtask. Following previous research, the goal is for the three ideas to be diverse [23]. However, here the ideas are chosen based on an underlying user model that is generated throughout the ideation session. Each time an inspiration is requested, the system will show the user one idea from each of the following categories: 1) an idea of the same category as the user's last generated idea; 2) an idea of a category that is adjacent to the user's current category; and 3) a new category that has not yet been visited by the user.

These categories have been selected to curb the two points of failure described in the SIAM model: failure to generate a new idea within the current category, and failure to retrieve a new category [18]. The first failure is addressed by showing the user their current category, attempting to inspire further ideas within it (effectively increasing fluency within the category). The second failure is addressed firstly by showing users an adjacent category, which is an idea category the user has transitioned to (from their current category) in the past. But it is possible that this adjacent category, which has been visited in the past, will not yield any new ideas. Furthermore, if the system is capable only of suggesting categories the user has visited in the past, it is possible that it would hinder the ideator's breadth by forcing their attention to those categories. Therefore, the system presents ideators with an idea within a category that has not yet been explored by the user. This idea, however, is not random, rather it is based on categories explored by similar users. In other words, the system acts as a recommendation system, suggesting new categories based on other similar ideators. This is explained in more detail in the next subsection (User Modeling).

The system also performs an ongoing **subtle adaptation** of the solution space by ordering its rows and columns. We call this ongoing reordering of the matrix. This reordering happens every time the user submits a new idea. The purpose for adapting this view is to guide users' attention to the most relevant categories. Since the goal for the solution space is to

give users an overview of all developed ideas, this adaptation can be seen as a form of tailoring example searching in a way that augments creativity [13]. The categories are ordered following the same logic as that of explicit adaptations. It orders the solution space, from right to left and top to bottom, in the following way: 1) current category; 2) all adjacent categories, ordered by most to least common; 3) *inferred new categories;* 4) other previously visited non-adjacent categories, sorted by most to least frequent; 5) any other category that has not yet been visited.

In comparison to the explicit interventions, this ongoing adaptation has the advantage of encoding more information, such as allowing users to explore overlaps between categories that are meaningful to them. It also retains user agency: rather than pushing three categories deemed useful to the user, they can choose what to explore in more detail. The downside, however, is that with more information being presented at once, ideators are less likely to pay attention to the ideas they encounter and therefore reduce the effect the ideas can have on them. This may be particularly meaningful when compared to the explicit inspirations, which employ microtasks to increase attention to ideas.

User Modeling

The adaptations described above are powered by an underlying user model. This model is inferred based on a user's behavior within the system. Whenever users add an idea, they are asked to choose one or two categories for their idea (see Figure 2). This selection is done through a list of existing categories, which the system uses to update the user's model. Based on the previous discussion on the adaptations supported by the system, the user model must be able to inform the system about four kinds of categories:

1) What is the user's current category? This is determined simply by looking at the last idea added by the user. The idea's category is considered to be the currently loaded category. If two categories were used, both are considered to be currently loaded.

2) From the current category, where is the user likely to move to? While the user ideates, the system keeps track of category transitions through a *transition graph*: a directed, weighted graph, in which each node represents a category. When the user adds an idea, the system creates an edge between the categories for the latest idea and the preceding ones. The weight of the edge increases as that transition is repeated. This is how the system determines the set of *adjacent categories*.

3) In which categories is the user most fluent? While the user ideates, the system also creates a *category vector* to keep track of the number of ideas the user adds for each category.

4) What are new categories the user has not yet visited but in which they are likely to be fluent? To infer this information, we draw from recommender system techniques [2]. The system uses the user's category vector to identify other ideators that have a similar ideation pattern to their own by calculating the correlation between the user's vector and other ideators'. We select the top five users with the highest correlation to make the inferences. Then, for every category the user being analyzed has not yet visited but that the other matched users have, the system calculates the average fluency. The category with the highest average is considered to be the one with the most potential for visitation.

STUDY 1

We evaluated the system through an online study on Prolific (www.prolific.ac). We limited participants to only those with 85% approval rating and above 18 years of age. We also prevented participation from anyone who took place in any pilots that we had run in the past. All data was collected across three one-day sessions, with two weeks between the first and last sessions.

In this study, we focus on evaluating whether the combination of the two adaptive features in CrowdMuse improves three well-recognized brainstorming metrics: fluency (number of ideas), breadth (how many categories are surveyed by one ideator), and depth (within category fluency) of ideation. We establish the following hypotheses:

H1. *An adaptive system will increase the number of ideas over a non-adaptive system*. By tailoring inspirations to categories that are more likely to be visited by an ideator, we expect them to be more effective at sparking that new idea that would otherwise have not been generated, either in a new category (thus increasing breadth) or in a previously visited category (thus increasing depth). The result from this is an increased overall number of ideas.

H2. *An adaptive system will increase the breadth of idea generation*. As postulated in H1, an adaptive system will increase overall fluency. We argue that an adaptive system will increase this breadth by showing users their inferred categories, that is, categories they have not yet visited but that are likely to be relevant to them based on similar ideators. In the SIAM model, this equates to delaying the failure in retrieving a new category [18].

H3. *An adaptive system will increase the depth of idea generation*. We hypothesize that by showing ideas in the current and adjacent categories, an adaptive system will increase the number of ideas suggested within each category, yielding an overall greater depth. In the SIAM model, showing the current or adjacent categories is an attempt to delay failure in the phase of idea generation within a given category or in another previously ideated category [18].

A secondary goal of the study was to more thoroughly test the system to determine how to improve it for future use. To do so, we individually evaluate each adaptation type (subtle and explicit) when testing our hypotheses.

Method

We posted a study request through the Prolific platform. Upon accepting the study, participants went through a short

tutorial explaining each part of the system individually. This tutorial also introduced the brainstorming problem with the following description: "*Prolific is a great website for researchers and participants alike. However, there is always room for improvement. Come up with as many ideas you can to improve Prolific in any way you can think of. Be as specific as possible in your ideas*". After completing the tutorial, a 15-minute timer would appear on top of the screen and start to count down. After the timer was done, a pop-up screen appeared with a link to a final questionnaire asking about demographics, their experience with the task, and perceptions of the system.

We used a between-subject 2x2 full factorial design with participants being randomly assigned to a combination of two factors:

1. Solution space (random/adaptive): rows and columns could be ordered randomly, or according to the user's model.
2. Inspiration mechanism (random/adaptive): inspirations could be selected randomly, or according to the user's model.

Idea Pool and Categorization

Since the adaptive mechanisms are powered by data from other users, we had to pre-populate the system with users and ideas for this first study. This data came from several pilots we ran on Prolific, which included 49 users and 189 ideas, organized across 54 categories. Categories were determined by the pilot ideators themselves, who could tag their ideas when adding them. One of the authors of this paper created the final list of categories from these user-generated categories, with redundant categories collapsed (e.g. the categories *chat*, *forum*, and *email* were all collapsed under *communication*). Thus, when selecting an adaptive inspiration, one of the 189 ideas would be presented based on the match between the desired inspiration category and its own category. New categories added by study participants during the study sessions were not visible to anyone other than the participant who added them. Therefore, all participants were exposed to the same set of categories throughout the duration of these studies.

Metrics

We primarily evaluate the effects on brainstorming performance through metrics related to breadth (how many idea categories a user has visited) and depth (the fluency within a given category). To do so, we use two metrics for each of these dimensions. The first metric comes from a manual categorization of the ideas generated, completed by two researchers (including one of the authors). At the beginning of this categorization, both researchers worked together to define the core categories. As the categorization progressed and no new categories started to appear, the researchers started to work independently, only occasionally discussing where some ideas should be assigned; 70% of ideas were categorized in this manner. The remaining ideas were categorized by a single researcher. Categorization was blind to the ideator's experimental condition. We then extracted **breadth** as the number of categories visited. **Depth** is defined as the largest number of ideas a user generated within a single category.

Additionally, following previous research, we calculated a metric based on Latent Semantic Analysis (LSA), using a corpus from [12] (comprised of ideas in a similar task on Amazon's Mechanical Turk, n=7199) as well as ideas from our pilots (n=591). Following the approach presented in [12], we built an ideation tree based on the similarity between ideas, in which each node is an idea attached to its most similar parent [12]. From this tree, breadth is derived as the number of children nodes of the root, and depth as the maximum number of nodes in one branch. We refer to these metrics as **tree breadth** and **tree depth**.

By using two different metrics, we avoid the bias introduced by a single type of metric. Each metric has its own tradeoffs. We expect the manual metrics to be more accurate, but they are very subjective—different people would likely come to different categorizations. The tree metrics, on the other hand, do not depend on subjective judgements, but they are likely more inaccurate, especially due to the origin of the corpus used to generate them. It should further be noted that both metrics are highly correlated to the user's fluency—someone who comes up with 20 ideas will very likely have higher breadth and depth numbers than someone who comes up with 2. Thus, in our analysis we control for fluency.

Results

In total, 115 Prolific users performed our study (42.6% female). Most participants described themselves as non-Hispanic White (75%), with the UK having the largest participation (25%). Participants were randomly assigned to conditions, but since some users quit the study before finishing it, the distribution across conditions was not perfectly balanced. We had 32 participants with neither adaptive mechanisms, 25 with only an adaptive inspiration mechanism, 28 with only an adaptive solution space, and 30 with both adaptive mechanisms.

H1: Fluency did not Change Across Factors

Our first hypothesis was that an adaptive system would increase the fluency by increasing both breadth and depth of ideation, since the adaptations would aid users in curbing idea generation failures in both loops of the SIAM model. To evaluate this hypothesis, we calculated a two-way ANOVA with fluency as outcome variable, and the presence of an adaptive solution space and the presence of an adaptive inspiration mechanism as fixed factors. We include the interaction between factors in the model. There was no significant main effect of the adaptive solution space, $F(1,111)=1.541$, $p=0.217$, or the adaptive inspiration mechanism, $F(1,111)=0.003$, $p=0.957$. We did, however, find a marginally significant interaction effect between factors on the fluency of participants, $F(1,111)=3.74$, $p=0.056$. Figure 3 demonstrates this interaction, along with the mean fluency values.

	Low Fluency		Average Fluency		High Fluency	
	Random	**Adaptive**	**Random**	**Adaptive**	**Random**	**Adaptive**
Breadth	2.55 (0.18)	2.77 (0.23)	4.72 (0.13)	4.98 (0.14)	6.88 (0.18)	7.20 (0.23)
Tree breadth	2.55 (0.24)	2.50 (0.30)	3.85 (0.17)*	4.66 (0.18)*	5.15 (0.24)**	6.82 (0.29)**

* $p < 0.05$; ** $p < 0.00$.

Table 1 Marginal means (and standard error) for breadth and tree breadth across different ideators of different fluency levels: low (1 sd below average), average, and high (1 sd above average).

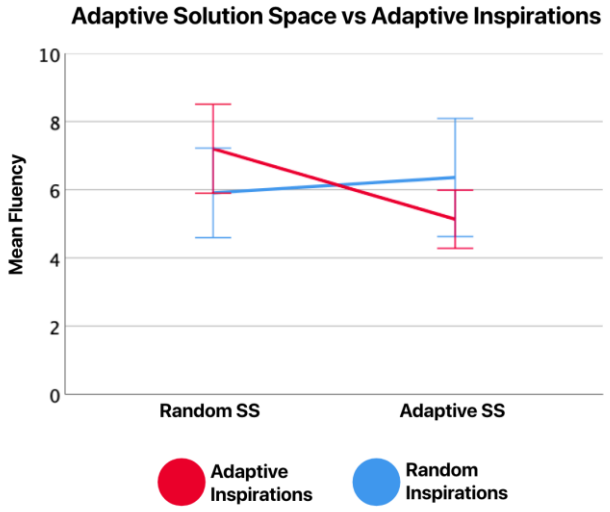

Figure 3 Interaction between an adaptive solution space and adaptive inspirations on fluency. Error bars represent a 95% confidence interval.

H2: Adaptive Inspirations Increased Tree Breadth

We first evaluated the manual breadth metric by running an ANCOVA with breadth as the outcome variable, adaptive solution space and adaptive inspirations as fixed factors, and fluency as a covariate. We found no effect of adaptive inspirations, $F(1,110)=2.721$, $p=0.102$, adaptive solution space, $F(1,110)=1.482$, $p=0.226$, or the interaction, $F(1,110)=0.358$, $p=0.551$.

We then evaluated tree breadth as an outcome variable. Because there was a significant interaction between fluency and one of our independent variables, we calculated a Mixed Generalized Linear Model (GLM), with the breadth metric as outcome variable, the presence of an adaptive solution space and an adaptive inspiration mechanism as fixed factors, and fluency as a covariate. We included two-way interactions between each factor and fluency. We found a significant interaction between adaptive inspirations and fluency, $F(1,108)=7.949$, $p = 0.006$, showing a stronger positive effect of adaptive inspirations on tree breadth as fluency increases. Pairwise comparisons show average and high fluency ideators who were exposed to adaptive inspirations outperformed those with randomized ones. Table 1 details the marginal means for the breadth metric across different fluency levels. We found no interactions between an adaptive solution space, $F(1,108)=0.706$, $p=0.402$, nor between the two main factors, $F(1,108)$, $p=0.286$. There were also no main effects of an adaptive solution space, $F(1,108)=0.187$, $p=0.666$, or an adaptive inspiration mechanism, $F(1,108)=0.308$, $p=0.308$.

H3: No Effects on Depth

Our third hypothesis was that by emphasizing ideas in their current or adjacent categories, ideators would likely be able to generate more ideas within those categories, therefore increasing depth. To evaluate this, we estimated a negative binomial regression, since the data follows a negative binomial distribution. We use depth as an outcome variable, presence of an adaptive inspiration mechanism as factor, and fluency as covariate. We found no significant effect on depth of either adaptive solution space, Wald Chi-Square=0.418, $p=0.518$, adaptive inspirations, Wald Chi-Square=0.211, $p=0.646$, or an interaction between both, Wald Chi-Square=0.007, $p=0.934$. The same model with tree depth as outcome variable equally yielded no effect from adaptive solution space, Wald Chi-Square=0.036, $p=0.850$, adaptive inspirations, Wald Chi-Square=0.294, $p=0.588$, or the interaction between the two factors, Wald Chi-Square=1.380, $p=0.240$. Therefore, the adaptations produced no change in depth.

Study 1 Discussion

In summary, we found some support for a positive effect of adaptive inspirations on the breadth of ideation (H2), although only from the tree metric. We did not find significant effects in either fluency (H1) or depth (H3), although we identified a marginal interaction on the former.

It is not surprising that the inspiration mechanism rather than the solution space was behind the change in breadth—it was designed to draw greater attention to the ideas (through the rating task), and to only show a reasonable amount of information per request. The question is why it only affected breadth, but not depth. One possible explanation is that the rating task may have pushed ideators to focus on originality. Eight ideators have said something to this effect. For example, one user reported that having to rate other ideas encouraged them "*to come up with original and feasible solutions*". Another ideator reported that rating other ideas "*allowed me to see what users said, and I started noticing patterns*" that allowed her to identify original ideas.

While we expected the inspiration mechanism to have greater effect, we did not think that the solution space would have no effect. This could perhaps be explained by issues in its usability and comprehension. Many users reported some issue with it such as finding it confusing, finding the tagging poor or redundant, or just having to scroll through so many tags (n=18). In fact, we find that perceptions of how useful the solution space was were correlated with perceptions of the usability of the system (ρ=0.467, p=0.000) and success in the task (ρ=0.216, p=0.022), which could mean that those who were confused by it had negative perceptions of the system or the task.

STUDY 2

We made the following changes to the system for Study 2:

- We recategorized the pool of ideas based on the manual categorization used in the breadth and depth metrics. In practice, this means that the solution space would be better organized, addressing complaints of poor, redundant, or excessive tagging. It should also reduce the noise in inspirations (e.g. miscategorized ideas), improving their effects. We have also included the ideas from Study 1 in the pool of ideas, meaning there are more ideas per category and more user models to draw from. In total, the system now had 899 ideas from 173 users, spread across 19 categories.
- The system now displays the ideas shown as inspirations differently—they are colored in blue and display a lightbulb icon. They are also always shown alongside the ideators' own ideas. We expect this to increase the effect of the inspirations by allowing users to examine them longer and more easily refine or combine them.
- We emphasized the refinement and combination actions. We do this in two ways. The first is by focusing on these actions in the introductory tutorial, explaining them more clearly. The second is by adding stats at the top of the idea workspace, mentioning how many original, refined, and combined ideas the user has added. We expect these changes to increase the usage of the refinement and combination actions.

Method

In this study, we only examined two conditions: control (inspirations and solution space are randomized) and fully adaptive (inspirations and solution space adapt to users). Our main hypotheses remain the same but due to the changes described above we expect a stronger effect in breadth, and possibly some effects in fluency and depth as well.

Study 2 Results

In total, 76 Prolific users participated in this study (40.8% female), 38 subjects in each condition, generating a total of 483 ideas. Most of them described themselves as non-Hispanic White (~61%), with the US (~22%) and the UK (~24%) having the highest number of participants.

We once again evaluated the same hypothesis from Study 1: H1: fluency will increase due to adaptations; H2: breadth will increase due to adaptations; and H3: depth with increase due to adaptations. We used the same statistical analyses employed in Study 1, with the difference being that now there is only one fixed factor, condition. Consequently, we only evaluated one interaction, between fluency and condition, in the GLMs.

H1: Fluency did not Change Across Conditions

Ideators in the control condition generated, on average, 6.29 ideas (sd=3.07). Those in the adaptive condition generated on average 6.39 ideas (sd=4.29). A One-way ANOVA shows no difference in fluency between conditions, F(1,74)=0.211, p=0.903.

H2: Adaptations Negatively Affected Tree Breadth

Fluency	Control	Adaptive	p
Low	1.88 (0.29)	2.39 (0.25)	0.190
Average	4.48 (0.19)	3.97 (0.19)	0.059
High	7.08 (0.29)	5.56 (0.24)	0.000

Table 2 Marginal means and standard errors for tree breadth in Study 2.

An ANCOVA showed no effect of condition on breadth, F(1,73)=1.280, p=0.262. Tree breadth, on the other hand, shows differences. A Mixed GLM shows a significant interaction between condition and fluency, F(1,72)=13.09, p=0.001. However, unlike Study 1, this time the interaction favored the control condition over the adaptive one, showing a marginal difference for average fluency ideators and a significant different for high fluency. Table 2 details how the marginal means change across low, average, and high fluency levels. There was also a main effect of condition, F(1,72)=5.052, p=0.028. Therefore, we find some evidence that an adaptive system hindered performance compared to a non-adaptive one.

H3: No Effects in Depth

We again found no effects of condition on either depth, Wald Chi-Square=0.009, p=0.925, or tree depth, Wald Chi-Square=0.034, p=0.854.

Study 2 Discussion

Like in study 1, we only found breadth effects. However, unlike the first study, this was a negative effect (on tree breadth only) caused by the adaptations. What could explain this difference? We did not find much difference in usage of the refinement and combination mechanics, so it is unlikely that this extra incentive is at its cause. We also find no reason for why the persistence of the inspirations in the workspace could have caused such an inversion of effect.

Therefore, we hypothesize that this striking change in effect was due to the new categorization. There was a dramatic reduction in the number of categories (54 in the first study to 19 in the second). It is possible that the number of categories was now too small to be meaningful, making them too high-level to be significant for both the adaptation and the metrics.

REVISITING THE CATEGORIZATION SCHEME

To examine the categorization effect on metrics, we calculated, for both studies, the number of distinct categories users were exposed to through the inspiration mechanism. We then estimated an ANCOVA with the number of categories as outcome variable, condition as fixed factor, and number of inspirations as covariate. For study 1, we find no effect of condition on the number of categories users were exposed to, $F(4,93)=0.535$, $p=0.66$. On average, study 1 participants across conditions have been exposed to 5.47 (sd=2.97) categories. Despite this, we still found a positive influence of an adaptive system on ideation breadth, indicating that that effect is not due to differences in quantity of categories, but rather on their quality. Study 2, on the other hand, showed a higher number of categories for the control condition (M=4.49, SE=0.19) compared to the fully adaptive one (M=3.85, SE=0.19), $F(2,73)=5.32$, $p=0.024$. A difference in the number of categories users were exposed to could partially explain the change in effect.

The remaining differences in effect could be due to the chosen ideas being less appropriate for each ideator. If the categories are too broad, and if there are more ideas per category, it is possible that choosing a random idea from within that category may not cause the desired inspiration effects. In fact, we find a marginal difference in how useful ideators in the control (5.82, sd=1.20) and adaptive (5.18, sd=1.66) conditions perceived the inspiration mechanism to be (on a 1-7 scale, 7 being the most useful), $F(1,74)=3.61$, $p=0.061$, indicating a trend towards greater dissatisfaction.

A categorization that is too high-level (i.e. too few categories) could also explain the lack of effects on the manual metrics in both studies, diluting nuanced category exploration. Therefore, we recalculated the manual metrics for both studies. One of the authors of this paper coded all ideas for both studies (N=1183) and developed a new categorization with 45 total categories. This new scheme increased the number of categories by breaking down the previous ones. For example, the original scheme had a category called *Study types*, which in the new categorization was broken down into categories such as *collaborative studies* or *in-person studies*. Another researcher then was given this new scheme along with 120 uncategorized ideas and independently categorized them. Agreement between raters was satisfactory, Cohen's Kappa=0.788.

With this new scheme, we revisited the analysis of the manual metrics in the previous two studies. In Study 1, we recalculated an ANCOVA with the new breadth metric as outcome variable, both adaptations as factors, and fluency as covariate. We found a main effect of adaptive inspirations on breadth, $F(1,110)=6.200$, $p=0.014$, with adaptive inspiration ideators exploring slightly more categories (M=6.61, SE=0.19) than those without the adaptive inspirations (M=5.96, SE=0.18). We still found no adaptive solution space effect, $F(1,110)=0.00$, $p=0.990$, as well as no interaction between the two factors $F(1,110)=0.528$, $p=0.569$. As for depth, we still found no effect of either adaptive inspirations, Wald Chi-Square=0.057, $p=0.812$, adaptive solution space, Wald Chi-Square=0.001, $p=0.976$, or the interaction between the two factors, Wald Chi-Square=0.126, $p=0.722$. These results reinforce those obtained through the tree metrics.

We also redid the analysis for Study 2. To evaluate breadth, we recalculated an ANCOVA with the new breadth as outcome, condition as factor, and fluency as covariate. This time we find no effect of condition on breadth, $F(1,73)=1.068$, $p=0.305$. We also re-evaluated depth, finding no significant effects, Wald Chi-Square=0.034, $p=0.854$.

DISCUSSION

From these two studies, we draw four conclusions. (1) Given an appropriate categorization scheme, adaptive inspirations can positively influence breadth of ideation. (2) Our adaptations, as they were proposed, were not capable of improving fluency or depth. (3) The inspiration mechanism had a stronger effect compared to the solution space. (4) The categorization scheme is key to the adaptations.

The inspiration mechanism's effect on breadth could be explained by the diversity of ideas presented. In study 2, where the variety of categories was decreased, we found evidence of the system performing as well or worse than control on breadth, potentially because ideas that were too similar were being presented to users. This finding is also in line with previous work, which found that diversity yields diversity [23]. But as we also found from a comparison of both studies, a difference in the total quantity of exposure categories does not completely explain this effect, as study 1 still revealed an advantage to adaptive inspirations despite an equivalent number of exposure categories. Therefore, we argue that with an appropriate categorization the adaptive inspiration mechanism is able to better select inspiration categories and improve breadth of idea generation.

In contrast, both studies showed a lack of significant effects on both fluency and depth. This suggests that the intended effects of the current and adjacent categories were not realized. Their intention was to keep users longer in the current categories, but for both studies we found considerably high likelihoods of users not staying within the same category for two consecutive ideas (95% on Study 1; 85% on Study 2). This lack of effect on depth likely contributed to the overall lack of effects on fluency. It may be that to be effective, the adaptation mechanism needs to better account for the fact that these ideators are likely to frequently switch categories.

We also note that the positive effect found in study 1 sprung from the adaptive inspirations, not the solution space. As we discussed in the system design, we expected that to be the case due to fewer ideas being presented at a time (compared to the solution space), as well as the built-in tasks. Both of these factors should increase the attention to the ideas that were presented, and therefore their effect. However, we also

acknowledge that in study 1 the solution space may have been plagued by usability issues, which may have distracted users from its benefits. We attempted to improve the usability of the solution space for study 2 by improving the categorization scheme, which brought its own set of issues.

Finally, a contrast between the studies points to the importance of the categorization scheme, both for powering the adaptations as well as for measuring their effects. The two studies showed markedly different results. We attribute that in great part to the reduced number of categories. Therefore, the right level of categorization abstraction is essential. The same applies for metrics. Our initial categorization was not detailed enough to capture differences between the factors, which was fixed by the later scheme.

It is worth further discussing the lack of effects in depth and fluency. For study 1, we had power to detect small effects on fluency (0.26) and breadth (0.08), and medium effects in depth (0.3). For study 2, we had power to detect small effects on breadth (0.1), medium effects on fluency (0.32) and large effects on depth (0.6). All of these were calculated considering 1-B=0.8 and our sample sizes. Our sample sizes are also in line with previous studies in the literature. Therefore, we argue that the null results do not stem from lack of power. Instead, we propose a different explanation based on past work and the SIAM model.

Research that used idea exposure as its inspiration mechanism has consistently found breadth improvements [12,23]. In contrast, inspirations based on simulation strategies (i.e. run a scenario in your mind) showed increases in depth and fluency, but not breadth [4]. The SIAM model points to a compelling conciliation of these results. It indicates that breadth improvements stem from the image loading loop, while depth improvements come from the idea generation loop. It seems, therefore, that idea exposure may be able to positively affect the image loading loop, but not the idea generation loop. On the other hand, another form of inspiration such as the simulation prompts described in Chan et al. [4] may be able to positively affect the idea generation loop, functioning as a form of mental exercise for exploring the currently loaded image in greater depth.

These results lead to implications for future work on CrowdMuse, suggesting the use of different types of inspirations for different types of categories. For example, instead of showing three ideas in the inspiration panel, it could show one (inferred) along with two simulation prompts related to the current and adjacent categories. The next step could be for CrowdMuse to adapt not only to idea categories, but also to cognitive strategies of individual ideators.

We conclude this discussion by addressing some limitations, beginning with the metrics. We built the tree metrics partly using a dataset of ideas from another crowd platform. Nonetheless, their content is rather similar to ours and it was augmented with ideas from our pilots. Furthermore, we complement these metrics with manually derived ones. Another metric limitation stems from our lack of evaluation of product metrics such as the originality and usefulness of the final ideas. Therefore, it is unknown whether the adaptations influenced the final product or not.

Another related, but broader, limitation stems from our use of discrete categories as the basis for the metrics and the system adaptations. This is a simplification of an ideator's cognitive structure, which is much more nuanced and complex [18]. However, given the complexity of accurately representing this structure, research often turns to categories as the basis for analysis (e.g. [3]), as we did in our studies. This limits the accuracy of the adaptations. Future work that adapts on finer-grained models is likely to yield better adaptations and, consequently, results.

There is also a concern with the high number of tests performed in each study, increasing the likelihood of type I errors. While we acknowledge this possibility, we note that all the tests performed were theoretically grounded, and the results are consistent with previous studies in the literature.

Finally, while we identified the sensitivity that the system has to the categorization scheme, there is still much more to be understood before reaching conclusions on best practices for categorization. Since we only compared two different schemes, there is not enough information to infer how the effects progress across a range of category numbers. Furthermore, there are also limited inferences we can make on the nature of categories themselves, rather than simply their numbers. In study 1, categories were user generated (with minor adjustments), while study 2 categories were generated by the researchers. The impact that this difference may have caused across both studies is unclear. These factors are key for CrowdMuse's usefulness in a real-world context, in which the categorization scheme would frequently change, especially at the earlier phases of idea generation, and therefore should be systematically evaluated in future work.

CONCLUSION

In this paper, we presented and evaluated CrowdMuse, a novel system that models and adapts to users to improve their ideation performance. We found that given an appropriate categorization, the adaptive inspirations were able to positively affect breadth of ideation. The adaptive solution space did not affect results, though issues of usability may have affected its effectiveness. Neither depth nor fluency were affected by adaptations. Finally, we also discussed the effect that categorization schemes of varying levels can have on the adaptations as well as measurements. We expect this work to open a new avenue for large-scale brainstorming support which can operate in synergy with other existing approaches to enhance the creative potential of crowds.

ACKNOWLEDGEMENTS

We would like to thank all Prolific users who participated in our studies, the reviewers for their thoughtful comments, as well as Shang Wang and Deniz Sonmez Unal for their help with data coding.

REFERENCES

[1] Bilal Alsallakh, Luana Micallef, Wolfgang Aigner, Helwig Hauser, Silvia Miksch, and Peter Rodgers. 2016. The State-of-the-Art of Set Visualization. *Computer Graphics Forum* 35, 1: 234–260. https://doi.org/10.1111/cgf.12722

[2] J. Bobadilla, F. Ortega, A. Hernando, and A. Gutiérrez. 2013. Recommender systems survey. *Knowledge-Based Systems* 46: 109–132. https://doi.org/10.1016/j.knosys.2013.03.012

[3] Vincent Brown, Michael Tumeo, Timothy S. Larey, and Paul B. Paulus. 1998. Modeling Cognitive Interactions During Group Brainstorming. *Small Group Research* 29, 4: 495–526.

[4] Joel Chan, Steven Dang, and Steven P. Dow. 2016. Improving Crowd Innovation with Expert Facilitation.

[5] Joel Chan, Steven Dang, and Steven P. Dow. 2016. Comparing Different Sensemaking Approaches for Large-Scale Ideation. Retrieved March 11, 2016 from http://joelchan.me/files/2016-chi-sensemaking-ideation.pdf

[6] Joel Chan and Christian D. Schunn. 2015. The importance of iteration in creative conceptual combination. *Cognition* 145: 104–115. https://doi.org/10.1016/j.cognition.2015.08.008

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

[8] Alan R. Dennis and Mike L. Williams. 2003. Electronic Brainstorming: Theory, Research, and Future Directions. In *Group creativity: Innovation through collaboration*. Oxford University Press.

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

[10] Karen Leggett Dugosh, Paul B. Paulus, Evelyn J. Roland, and Huei-Chuan Yang. 2000. Cognitive stimulation in brainstorming. *Journal of Personality and Social Psychology* 79, 5: 722–735. https://doi.org/10.1037/0022-3514.79.5.722

[11] Gerhard Fischer. 2005. Distances and diversity: sources for social creativity. 128. https://doi.org/10.1145/1056224.1056243

[12] Victor Girotto, Erin Walker, and Winslow Burleson. 2017. The Effect of Peripheral Micro-tasks on Crowd Ideation. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*.

[13] Scarlett R. Herring, Chia-Chen Chang, Jesse Krantzler, and Brian P. Bailey. 2009. Getting inspired!: understanding how and why examples are used in creative design practice. In *Proceedings of the 27th international conference on Human factors in computing systems - CHI 09*, 87. https://doi.org/10.1145/1518701.1518717

[14] Gaoping Huang and Alexander J. Quinn. 2017. BlueSky: Crowd-Powered Uniform Sampling of Idea Spaces. 119–130. https://doi.org/10.1145/3059454.3059481

[15] D. G. Jansson and S. M. Smith. 1991. Design Fixation. *Design Studies* 12, 1: 3–11.

[16] Nicholas W. Kohn, Paul B. Paulus, and YunHee Choi. 2011. Building on the ideas of others: An examination of the idea combination process. *Journal of Experimental Social Psychology* 47, 3: 554–561. https://doi.org/10.1016/j.jesp.2011.01.004

[17] Bernard A. Nijstad, Michael Diehl, and Wolfgang Stroebe. 2003. Cognitive Stimulation and Interference in Idea-Generating Groups. In *Group Creativity: Innovation Through Collaboration*. Oxford University Press.

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

[19] Bernard A. Nijstad, Wolfgang Stroebe, and Hein FM Lodewijkx. 2002. Cognitive stimulation and interference in groups: Exposure effects in an idea generation task. *Journal of experimental social psychology* 38, 6: 535–544.

[20] Alex F. Osborn. 1963. *Applied imagination; principles and procedures of creative problem-solving*. Scribner, New York.

[21] Paul B. Paulus and Vincent R. Brown. 2003. Enhancing ideational creativity in groups: Lessons from research on brainstorming. In *Group creativity: Innovation through collaboration*. Oxford University Press.

[22] Matti K Perttula and Lassi A Liikkanen. 2006. Exposure Effects in Design Idea Generation: Unconscious Conformity or a Product of Sampling Probability? 14.

[23] Pao Siangliulue, Kenneth C. Arnold, Krzysztof Z. Gajos, and Steven P. Dow. 2015. Toward Collaborative Ideation at Scale: Leveraging Ideas from Others to Generate More Creative and Diverse Ideas. 937–945. https://doi.org/10.1145/2675133.2675239

[24] Pao Siangliulue, Joel Chan, Steven P. Dow, and Krzysztof Z. Gajos. 2016. IdeaHound: Improving Large-scale Collaborative Ideation with Crowd-Powered Real-time Semantic Modeling. 609–624. https://doi.org/10.1145/2984511.2984578

[25] Pao Siangliulue, Joel Chan, Krzysztof Z. Gajos, and Steven P. Dow. 2015. Providing Timely Examples Improves the Quantity and Quality of Generated Ideas. 83–92. https://doi.org/10.1145/2757226.2757230

[26] Thomas B. Ward and Yuliya Kolomyts. 2010. Cognition and Creativity. In *Cambridge Handbook of Creativity*. 93–112.

[27] Lixiu Yu, Aniket Kittur, and Robert E. Kraut. 2014. Distributed analogical idea generation: inventing with crowds. 1245–1254. https://doi.org/10.1145/2556288.2557371

[28] Lixiu Yu, Aniket Kittur, and Robert E. Kraut. 2014. Searching for analogical ideas with crowds. 1225–1234. https://doi.org/10.1145/2556288.2557378

[29] Lixiu Yu, Aniket Kittur, and Robert E Kraut. 2016. Encouraging "Outside- the- box" Thinking in Crowd Innovation Through Identifying Domains of Expertise. 1212–1220. https://doi.org/10.1145/2818048.2820025

[30] Lixiu Yu, Robert E Kraut, and Aniket Kittur. 2016. Distributed Analogical Idea Generation with Multiple Constraints. 1234–1243. https://doi.org/10.1145/2818048.2835201

[31] Lixiu Yu and Jeffrey V. Nickerson. 2011. Cooks or cobblers?: crowd creativity through combination. In *Proceedings of the SIGCHI conference on human factors in computing systems*, 1393–1402. Retrieved October 12, 2015 from http://dl.acm.org/citation.cfm?id=1979147