Proceedings of the ASME 2018 International Design Engineering Technical Conferences & Computers and Information in Engineering Conference
IDETC/DTM 2018
August 26-29, 2018, Quebec City, Quebec, Canada

IDETC2018-85470

UNPACKING SUBJECTIVE CREATIVITY RATINGS: USING EMBEDDINGS TO EXPLAIN AND MEASURE IDEA NOVELTY

Faez Ahmed*
Dept. of Mechanical Engineering
University of Maryland
College Park, Maryland 20742
Email: faez00@umd.edu

Mark Fuge
Dept. of Mechanical Engineering
University of Maryland
College Park, Maryland 20742
Email: fuge@umd.edu

Sam Hunter
Industrial and Organizational Psychology
The Pennsylvania State University
University Park, PA
Email: sth11@psu.edu

Scarlett Miller
School of Engineering Design, Technology and Professional Programs
The Pennsylvania State University
University Park, PA
Email: shm13@psu.edu

ABSTRACT

Assessing similarity between design ideas is an inherent part of many design evaluations to measure novelty. In such evaluation tasks, humans excel at making mental connections among diverse knowledge sets and scoring ideas on their uniqueness. However, their decisions on novelty are often subjective and difficult to explain. In this paper, we demonstrate a way to uncover human judgment of design idea similarity using two dimensional idea maps. We derive these maps by asking humans for simple similarity comparisons of the form "Is idea A more similar to idea B or to idea C?" We show that these maps give insight into the relationships between ideas and help understand the domain. We also propose that the novelty of ideas can be estimated by measuring how far items are on these maps. We demonstrate our methodology through the experimental evaluations on two datasets of colored polygons (known answer) and milk frothers (unknown answer) sketches. We show that these maps shed light on factors considered by raters in judging idea similarity. We also show how maps change when less data is available or false/noisy ratings are provided. This method provides a new direction of research into deriving ground truth novelty metrics by combining human judgments and computational methods.

INTRODUCTION

Creativity is the driving force of innovation in design industry. Despite many methods to help designers enhance creativity of generated ideas, not much research has focused on what happens after the generation stage [1]. One of the main problems that design managers face after completion of an ideation exercise is that of idea evaluation. Contributors have just sent in a flood of design ideas of variable quality, and these ideas must now be reviewed, in order to select the most promising among them. Idea evaluation has been highlighted as a central stage in the innovation process in fields like design and management [2]. However, many of the existing methods in idea evaluation using creativity metrics are quite subjective. An emerging thread of research within idea evaluation is on attempts to quantitatively assess creativity of ideas [3–5]. Here, creativity of ideas is often

*Address all correspondence to this author.

viewed as the comparison of design options on key factors like novelty and quality. Novelty is generally understood as uniqueness of an idea or how different it is from everyone else [6], while quality can be defined using multiple domain dependent factors like functionality, feasibility, usefulness, impact, investment potential, scalability *etc.* To distinguish this from novelty, quality is a measure of the designs' performance rather than a measure of how it differs from other designs in its class [7].

Novelty of ideas can be measured by experts or non-experts. Experts have a substantial knowledge of the field and of the market, and can thus provide more informed and trustworthy evaluations [8]. Many crowdsourcing platforms such as Topcoder, Taskcn, and Wooshii use expert panels to select contest winners [9]. However, experts are also scarce and expensive, since gaining expertise on a particular innovation subfield takes a substantial amount of training. Crowds have been proposed as an alternative [10] to evaluating ideas due to their large diversity of viewpoints, knowledge and skills (*i.e.*, the "wisdom of the crowds" [11]).

Ideas are often judged on novelty by crowds or experts, which gives rise to two important research issues in idea evaluation. First, what scale should be used by people to judge novelty of ideas? Second, how can one explain the decision making process of idea evaluators? In this paper, we try to answer the second question by calculating what we call *idea maps—i.e.*, an embedding or mapping of ideas into an N-dimensional Euclidean space—for raters and then estimating the novelty of those ideas from those maps.

RELATED WORK

In this section, we review research related to creativity ratings and design space visualization and how the two can be combined to help create explainable metrics.

Creativity Ratings

In the social sciences, creativity is often measured subjectively through the Consensual Assessment Technique (CAT) [12]. They define a creative idea as something that experts in the idea's or project's focus area independently agree is creative. However, it is difficult to explain what factors are used by experts to give a particular novelty score to ideas. As humans have limited memory, it is also possible that while judging novelty of every idea, experts may not remember all existing ideas similar to it or they underestimate the originality of truly novel ideas [13]. By using different attributes or different criteria of evaluation within the same attribute, it is possible that experts decide on completely different "novel" items.

In contrast, engineering design creativity research focuses on the measurable aspects of an idea by breaking down the concept into its different components and measuring their creativity in various ways [5, 14, 15]. Despite existence of multiple metrics in engineering design for measuring design creativity, most methods have been heavily criticized for their lack of generalizability across domains, the subjectivity of the measurements and the timeliness of the method for evaluating numerous concepts [16, 17]. For example, one of the commonly used tree-based metrics—SVS [14]—breaks down creativity into quantity, quality, novelty, and variety. The resultant novelty score of an idea depends on which attributes are considered in the tree and may vary between two different raters/trees [18].

Design Space Visualization

One way to better understand the decision making of raters is to visualize the design space by placing all ideas on a map and grouping similar items together. Design space exploration techniques [19] have been developed to visualize a design space and generate feasible designs. Motivated by the fact that humans essentially think in two or three dimensions, many methods to visualize high dimensional data by mapping it to lower dimension manifolds have been studied extensively [20, 21]. Researchers have also investigated information extraction about the broader nature of a design space from such embeddings [22]. In a typical machine learning setting, one assumes to be given a set of items together with a similarity or distance function quantifying how "close" items are to each other. A difficulty in creating low dimensional manifolds for design ideas is that complex design ideas often lack compact vector representations or known similarity measures. One solution to this problem is to directly ask humans about how similar ideas are.

There are two common ways to collect similarity ratings from people. In the first way, one typically asks people to rate the perceived similarity between pairs of stimuli using numbers on a specified numerical scale (such as a Likert scale) [23]. Methods like classical multi-dimensional scaling [24] can be used with these ratings to find an embedding. However, these ratings are not considered suitable for human similarity judgments as different raters use different "internal scales" and raters may be inconsistent in their grading [25].

As humans are better at comparing items than giving absolute scores [26], the second way is to gather ordinal judgments. For instance, triplet ratings consists of asking subjects to choose which pair of stimuli out of three is the most similar in the form "Is A more similar to B or to C?". Once similarity judgments are captured, one can use a number of machine-learning techniques that try to find an embedding that maximally satisfies those triplets and facilitate the visual exploration. Examples of such techniques include Generalized Non-metric Multidimensional Scaling (GNMDS) [27], Crowd Kernel Learning [28] and Stochastic Triplet Embedding [25]. Such methods take in triplet ratings and output either an embedding or similarity kernel between items which best satisfy human triplet responses.

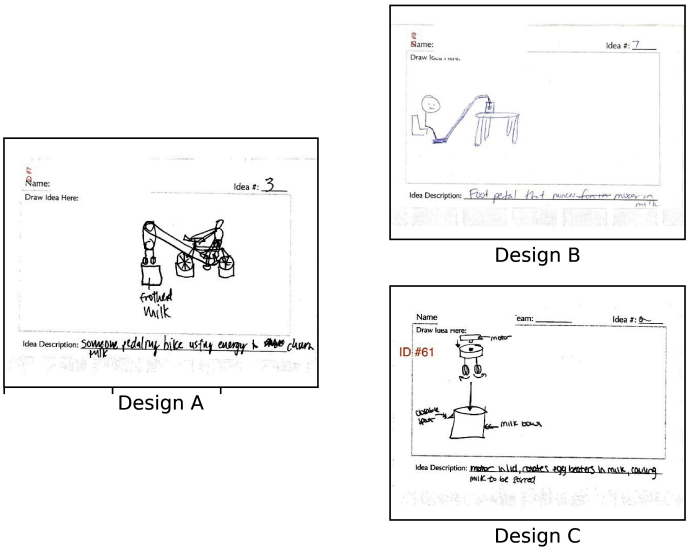

FIGURE 1. Example of sketch triplet used in experiment 2. Rater answers the question: "Which design is more similar to design A?

Techniques for capturing similarity among items using triplets have been applied in many areas like computer vision [29], sensor localization [30], nearest neighbor search [31] and density estimation [32]. In [33], authors learn perceptual kernels using different similarity methods. They find that triplet matching exhibits the lowest variance in estimates and is the most robust across the number of subjects compared to pairwise Likert rating and direct spatial arrangement methods. Our work is most similar to Siangliulue *et al.* [34], who use triplet similarity comparisons by crowdworkers to create spatial idea maps. They show that human raters agree with the estimates of dissimilarity derived from idea maps and use those maps to generate diverse idea sets. Our work differs from their work in two ways. First, we use idea maps to measure novelty of design ideas. Second, we try to uncover attributes factoring into how a rater decides the novelty of design sketches.

In this paper we focus on learning idea maps or design embeddings, *i.e.*, an embedding in which similar ideas lie close together and dissimilar ideas are far apart, entirely based on the similarity-triplets supervision provided by a person. We show how studying idea maps allows us to understand what factors may be important for different individuals in judging similarity and how these embeddings can be used to rate ideas on novelty. The next section provides an overview of the methodology used, followed by our experimental results on two design domains. We discuss the limitations and design implications, followed by various ideas to extend this method to derive new novelty metrics.

METHODOLOGY

In this section, we discuss the embedding method used and define a novelty metric based on idea maps.

Idea Map Generation

We first generate all possible triplets from a collection of N design ideas and collect responses on all these triplets from a group of raters. Then we use the Generalized Non-metric Multidimensional Scaling (GNMDS) technique [27] to find embeddings of design ideas.[1] GNMDS aims to find a low-rank kernel matrix K in such a way that the pairwise distances between the embedding of the objects in the Reproducing Kernel Hilbert Space (RKHS) satisfy the triplet constraints with a large margin. It minimizes the trace-norm of the kernel in order to approximately minimize its rank, which leads to a convex minimization problem. Figure 1 shows an example of a triplet with three design sketches used in our study. Triplet responses are represented as 'ABC' or 'ACB'. 'ABC' means Design A is closer to Design B than Design C and 'ACB' means Design A is closer to Design C than Design B. GNMDS allows the triplets to contradict; this can often happen when multiple people vote and use different criteria in finding item similarity. The resulting output is x,y coordinates for each design item. We then scale the map such that coordinates are between 0 and 1.

Measuring Novelty on a Map

Assuming we have obtained an idea map by applying an embedding method to the triplet responses by a rater, which satisfy a majority of consistent triplets, our next task is to calculate which ideas are novel in this map. As nearby ideas on the map denote similarity with each other, one would expect that the idea furthest away from everyone else will also be the most unique to the set. As novelty of an item in a set can be interpreted as how unique or dissimilar an item is [6], the problem is equivalent to finding ideas which are distant from all other ideas on the map. Hence, we define a metric which gives a high score to ideas which are away from everyone else on a 2-D map and a low score to items surrounded by many other ideas. To quantify this, the novelty score of item i is defined as:

$$Novelty\ (i) = \sum_{j=1}^{N} d_{i,j} \tag{1}$$

Here $d_{i,j}$ is distance of idea i from idea j in the two dimensional embedding. Hence, the novelty of an idea in a set is equal to the sum of distances from the idea to all other ideas. This

[1] Before selecting GNMDS, we compared it to three other common techniques—Crowd Kernel Learning, Stochastic Triplet Embedding and t-Distributed Stochastic Triplet Embedding—for our data. We did not find major differences in percentage of triplets satisfied between different methods.

simple formulation has been used in the past for document summarization to define representative items [35] and it allows us to score and rank order all ideas by novelty. We experimented with a few other methods to measure novelty of items on a map, but chose this metric as it is easy to compute and does not make assumptions about the distribution of ideas on the map.

Measuring Rater Performance

So far we have assumed that triplet responses are given by a rater, but it is important to consider that different raters may provide triplets of different quality. It is difficult to assess the quality of triplets, as they are essentially subjective assessment of how a rater views the similarity of ideas. Despite this, we estimate a rater's performance by measuring how consistent they are with their own responses using two methods. First, we estimate the self consistency of raters by adding additional triplets, which are repeats of the existing triplets. Second, we measure the number of violations a rater makes in the transitive property of inequality; this means, suppose a rater gives two responses as *ABC* and *CAB*. This means he finds item A more similar to item B and item C more similar to item A. These answers imply $AB < AC$ and $CA < CB$ for two different triplet queries, where *AB* denotes distance between idea A and idea B. These two inequalities imply that $BA < BC$, that is, idea B is more similar to idea A. If this rater provides a third triplet as *BCA* with idea B being more similar to idea C, then this violates transitive property—any two triplets are consistent, but not all three, so there is one violation of the transitive property.

We measure the total number of violations and the percentage of self consistent answers as measures of rater performance. In this study, we do not use explicit criteria to filter out raters with lower scores on either metric but this information could be incorporated in future studies to give higher weight to triplets of raters who are more self-consistent.

EXPERIMENTAL RESULTS

To demonstrate our methodology, we consider two case studies. We choose the first case study, such that the idea maps generated are simple to understand and the novelty measure is easily verifiable. By selecting items with only a few attributes, we can estimate the ground truth of novelty estimation and use it to verify our methodology. In contrast, for the second case study, we select a complex design domain, where "ground truth" is not known and different raters may disagree on what defines being novel. With this guiding principle, in the first study, we generate a dataset with ten colored polygons, who are rated by eleven raters. We show two dimensional idea maps and novel items discovered for different raters in a seemingly simple design domain. In the second study, we selected ten milk frother sketches from a real-world ideation exercise conducted as part of a previous paper [1]. Here we show how individuals vary in defining similarities between complex designs and how their ratings can be aggregated to generate meaningful idea maps.

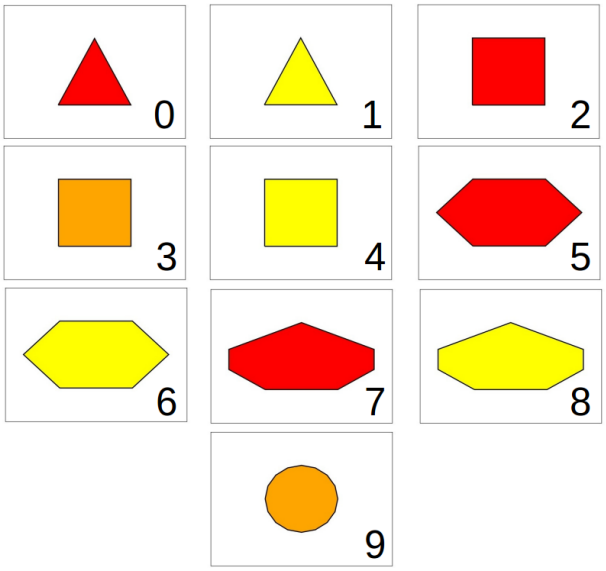

FIGURE 2. Dataset of ten polygons used in experiment 1

Experiment 1: Colored Polygons

Our dataset of ten polygons is shown in Fig. 2, which contains two triangles, three squares, two hexagons, two heptagons and one circle. We obtain 360 triplet queries (all possible permutations of three items) from these ten sketches and show them to eleven raters. The raters comprised one Ph.D. student (Industrial Engineering), one Master's student (Mechanical Engineering) and nine under-graduates (Psychology). Suppose a given triplet has items A, B and C as polygons 7, 6 and 2 from Fig. 2 respectively. For this triplet, raters decide whether they find the red heptagon more similar to the yellow hexagon or the red square. One rater may believe that color-based similarity is more important than shape and thus answer "the red heptagon is more similar to the red square," while another might believe the red heptagon is more similar to the yellow hexagon due to their closeness in area. To gain insights into their decision making process, we also asked raters to explain their choice for 20 randomly selected triplets. These responses helped verify our hypotheses about the factors considered by a rater compared to observing their map.

Automated rater To check if the triplet generated maps reflect the responses in ratings, we first use an automated rater who rates all triplet queries consistently based on fixed rules. We define the rules such that this automated rater always rates polygons with least difference in number of sides as more similar. Second, when both polygon B and C have similar priority in previous rule, it selects the polygon which is more similar in color. As

the automated rater uses consistent rules for all triplets, we find that its self consistency score is 100% and it has zero transitive violations. The resultant idea map obtained from the automated raters triplet ratings is shown in Fig. 3. One can notice from this idea map that similarly shaped items are grouped together. As one might expect, the two dimensions that can be identified from this idea map are color and shape. Polygons of similar shape are grouped together, while yellow colored polygons are placed slightly below their red counterparts. The gap between triangles and squares is lesser compared to the gap between squares and hexagons. This is because triplets with less difference in their number of sides were rated as more similar by the automated rater. Hence, we verify that this map is a good representation of the ratings provided by the rater.

In contrast to the automated rater, human raters may not always use consistent rules. Different people may give different priority to polygon attributes like color, shape, symmetry *etc.*. Table 1 shows the scores of self consistency and transitive violations, along with the most novel items calculated using Eq. 1 for all 11 raters. We also report the percentage of triplets not satisfied by the map.

Let us take the example of idea maps obtained for two raters (rater id 5 and rater id 9 from Table 1 respectively). One can notice from the map for rater 5 in Fig. 4, similar shaped polygons are placed near to each other. We also notice that red colored polygons are placed above yellow ones, similar to the automated rater. This provides evidence that this rater used shape and color as main criteria for decision making. In contrast, the map for rater 9 in Fig. 5 indicates that shape may not be as important as color for this rater. This is evident by the change in position of the orange square, which is closer to orange circle of similar color and far from similarly shaped squares.

When we look at the explanation provided by rater 5 for some of the triplets, she repeatedly mentions "My choice was made by determining which polygon had a number of sides closest to polygon A" while rater 9 mentions many of her triplet comparisons were decided by "color, shape, number of sides" as the reason. Hence, the criteria used by individual raters are reflected in their idea maps, grouping similarly colored or shaped items together.

Next, we focus on finding the most novel item. Intuitively, one would expect the most novel item to be most dissimilar to everyone else. One heuristic for verifying this finding can be counting the sketch which appears the most on A, B, C triplet ratings at position 3. This implies that a particular polygon is dissimilar in most triplet queries and is expected to be novel. For rater 5, we find that Sketch 9 (circle) occurs only 4 times. This is followed by yellow triangle (21 times) and then red triangle (24 times). We can also visualize the same using the map, where these polygons are on the periphery of rater 5's idea map.

Figure 4 shows that the circle is far away from all other polygons and thus one may consider it novel with respect to other polygons present in the dataset. To quantify this, we calculate the novelty score of a polygon as the sum of the distances with all other polygons for every rater. Table 1 shows the top three most novel sketches for each rater. We find that the orange circle appears in top three for most raters, indicating the novelty metric successfully captures the circle as the most novel item and there is consensus among raters that it is the most novel item in the set. This matches our expectations, as we generated this dataset such that the circle is of a different color and unique shape compared to other items in the set. Hence, by studying individual idea maps and calculating novelty measure of items on these maps, we can calculate the most novel items as well as understand the factors which different people consider in deciding item similarity.

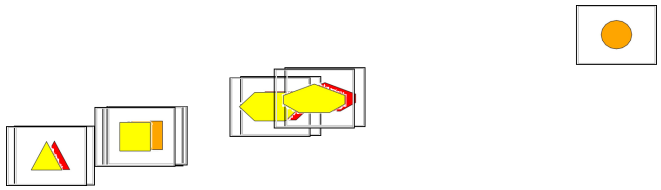

FIGURE 3. Two dimensional embedding for automated rater for polygons

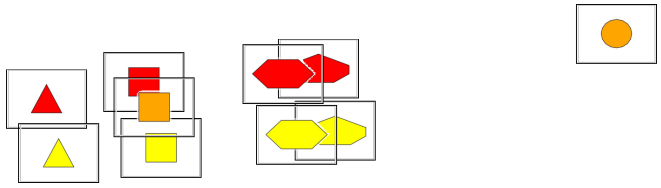

FIGURE 4. Two dimensional embedding obtained from polygon dataset by rater 5, who uses number of sides as primary criteria for triplet decisions

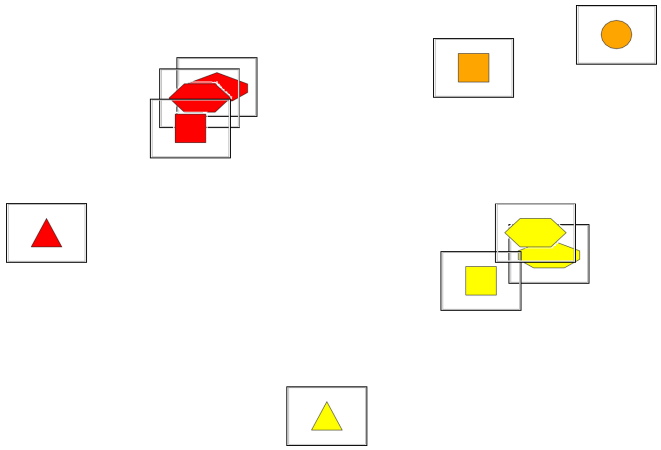

FIGURE 5. Two dimensional embedding obtained from polygon dataset by rater 9, who uses 'color, shape, number of side' as criteria

Rater ID	Self Consistency	Top 3 Most Novel Polygons	Number of Transitive Violations	Percentage not satisfied
AR	100.0%	9, 1, 0	0	5
1	83.3%	9, 1, 0	2	11
2	100.0%	9, 0, 8	3	11
3	83.3%	9, 4, 8	3	15
4	75.0%	9, 1, 0	2	15
5	100.0%	9, 1, 0	0	1
6	100.0%	9, 1, 0	0	15
7	91.6%	9, 1, 0	0	15
8	91.6%	9, 1, 6	8	21
9	83.3%	1, 9, 0	9	22
10	83.3%	9, 1, 3	4	10
11	100.0%	9, 8, 1	0	15

TABLE 1. Rater performance and Top 3 novel items for different raters of experiment 1 on polygons. We find that most raters find circle (item 9) as the most novel polygon.

Experiment 2: Design Sketches

In this experiment, we find the embeddings for ten design sketches of milk frothers. This set of design sketches is adopted from a larger dataset of milk frother sketches [1, 36]. To create the original dataset, the authors recruited engineering students in same first-year introduction to engineering design course. The task provided to the students was as follows: *"Your task is to develop concepts for a new, innovative, product that can froth milk in a short amount of time. This product should be able to be used by the consumer with minimal instruction. Focus on developing ideas relating to both the form and function of the product"*. Details of experiment to collect data are available online.[2]

We selected ten design sketches from this dataset for this experiment. Fig. 6 shows these sketches. As shrinking the sketches and their overlap makes it difficult to understand a 2-D map, we allocate number ids to each sketch and plot the numbers on idea maps instead. Similar to the previous case, eleven raters were used in this experiment. The raters comprised of one professor (Industrial Engineering), two Ph.D. students (Industrial Engineering) and seven under-graduate students (Psychology).

Figure 7 and 8 show the idea maps obtained by raters 7 and 10. These maps provide useful cues into the decision making process of these raters, who used different decision making criteria. The embedding of rater 7 in Fig. 7 provides evidence that she might have grouped sketches which have cup to store milk in the design as more similar (as shown by sketches 6, 5, 2 and 7). She also grouped sketches 4 and 3 nearby, both of which have bikes in the design. Similarly, rater 10 also has sketches 4 and 3 nearby but 6, 5, 2 and 7 are not nearby. To understand the rationale used by the two raters, we looked at their explanation for the triplet query shown in Fig. 1. Rater 7 finds sketch C as more similar to sketch A and mentions her choice as being based on "Simple or complex" design. Rater 10 finds sketch B as more similar to sketch A and gives the reason "it both spins and is powered by a person." We find rater 7 mentions for many other triplet queries that she used design complexity as the primary criteria for judging which ideas are similar. She also gives the reason: "If it spins, or if it includes cups" for a few triplets, indicating that the presence of cup is an important criteria in her decision making.

In contrast, rater 10, mentions a multitude of factors for different triplets like the method by which the milk was frothed (e.g. shaking), the form of the frother, if design had a motor, if something is being put into the milk or if the milk goes into something, *etc.* Due to the multitude of factors used by rater 10, ideas in her map may be positioned due to a combination of different factors. To verify the novelty calculation for rater 10, we asked her to provide us a rank ordered list of the most novel milk frother sketches from this dataset, without showing her the idea map generated by her triplets. Her top three most novel sketches were $0, 1$ & 6, while our novelty metric finds sketches $4, 0$ & 1 as the top 3 ideas from her idea map. This indicates that novelty calculation from her idea map is partially capable of capturing her assessment of novelty. It should also be noted that all three of her top 3 sketches $(0, 1, 6)$ occur on the periphery of her idea map (as shown in Fig. 8), showing that they are generally far away from other sketches.

We also found differences between justifications given by expert raters (experts are identified as raters with significant experience in rating milk frothers) and novices, where the latter focused more on surface level similarities while experts considered deep functional features too. In future work, we plan to study the differences between expert and novice idea maps.

Wisdom of the crowd Table 2 shows the self consistency score, transitive violations and top three most novel sketches for all users. As expected, maps of different raters differed from each other, which led to most novel ideas calculated using Eq. 1 differing too. As one would expect, we noticed that self consistency scores and transitive violations are larger for design sketches compared to polygons experiment, implying that it is more difficult to judge real-world sketches compared to polygons.

To understand how sketches are grouped together by the group of raters, we combine the triplet responses of all raters and obtain a joint idea map. Fig. 9 shows the joint map of all

[2] http://www.engr.psu.edu/britelab/resources.html

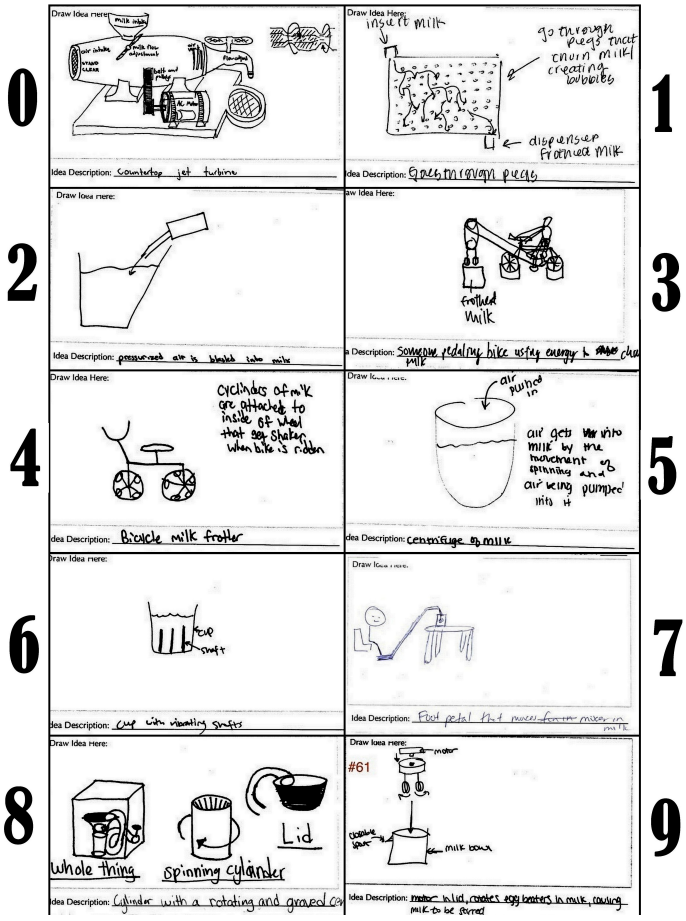

FIGURE 6. Ten milk frother sketches used in experimental 2

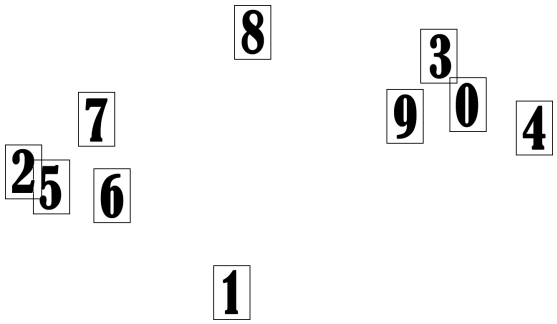

FIGURE 7. Idea map of design sketches for rater 7

eleven raters. As we add all triplets from raters who considered different (unknown) factors in judging idea similarity, the aggregated map can be considered to represent an average of all such attributes. One can study this map to find meaningful clusters in it and see which ideas are grouped together. For instance, on the right-hand side, we see three sketches (sketch 2, 5 and 6) clustered together, each of which uses a cup to hold milk. On the left-hand side, we see two sketches with bikes (sketch 3 and 4)

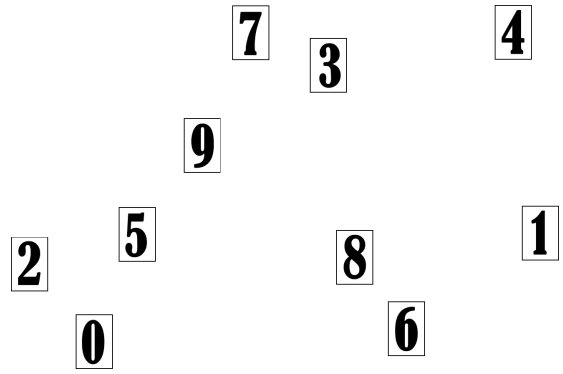

FIGURE 8. Idea map of design sketches for rater 10

clustered together. Two complex designs (sketch 0 and 8) with multiple moving parts are clustered together at the bottom. Using this map and our novelty metric, we find the most novel idea is sketch 0, while the least novel is sketch 9. Sketch 0 is at the bottom of the map in Fig. 9, quite distant from all other sketches. As noted before, sketch 0 proposing a counter-top jet turbine to froth milk is the most novel sketch rated by the expert too. While individual idea maps of different raters disagreed on scoring the most novel sketch (due to different criteria used), we also found that Sketch 9 ranked among the least novel items by majority of the raters.

So far, we have shown how individual idea maps can provide cues into factors important for raters in judging idea similarity. We have also shown how a joint map obtained by combining triplets from multiple raters can position sketches in a meaningful way and can be used to estimate explainable novelty of sketches. Next, we measure how raters differ from each other in their triplet responses.

Similarity between raters To compare the similarity between triplet responses of different raters, we represented their responses as a one-hot encoded binary vector of length 720 and found cosine similarity between these vectors.

We applied several clustering (*e.g.*spectral clustering) to these vectors and identified two clusters. The clustering analysis showed that raters 1, 3, 5 and 10 are in first cluster and all other raters are in second cluster. Interestingly, rater 5 and 10 were the two experts in our rater pool and we found that they were also clustered together, along with rater 1 and rater 3. We then calculated the similarity matrices for each user's idea map and found the matrix distance between different idea maps. We again clustered the raters using the distance between their maps, and found that they likewise group into two clusters. This finding is important, as we are able to find two supposedly non-experts, who are indistinguishable from experts based on their triplet ratings. Such groupings can be used to find aggregated maps for each group and study differences between idea maps for a group

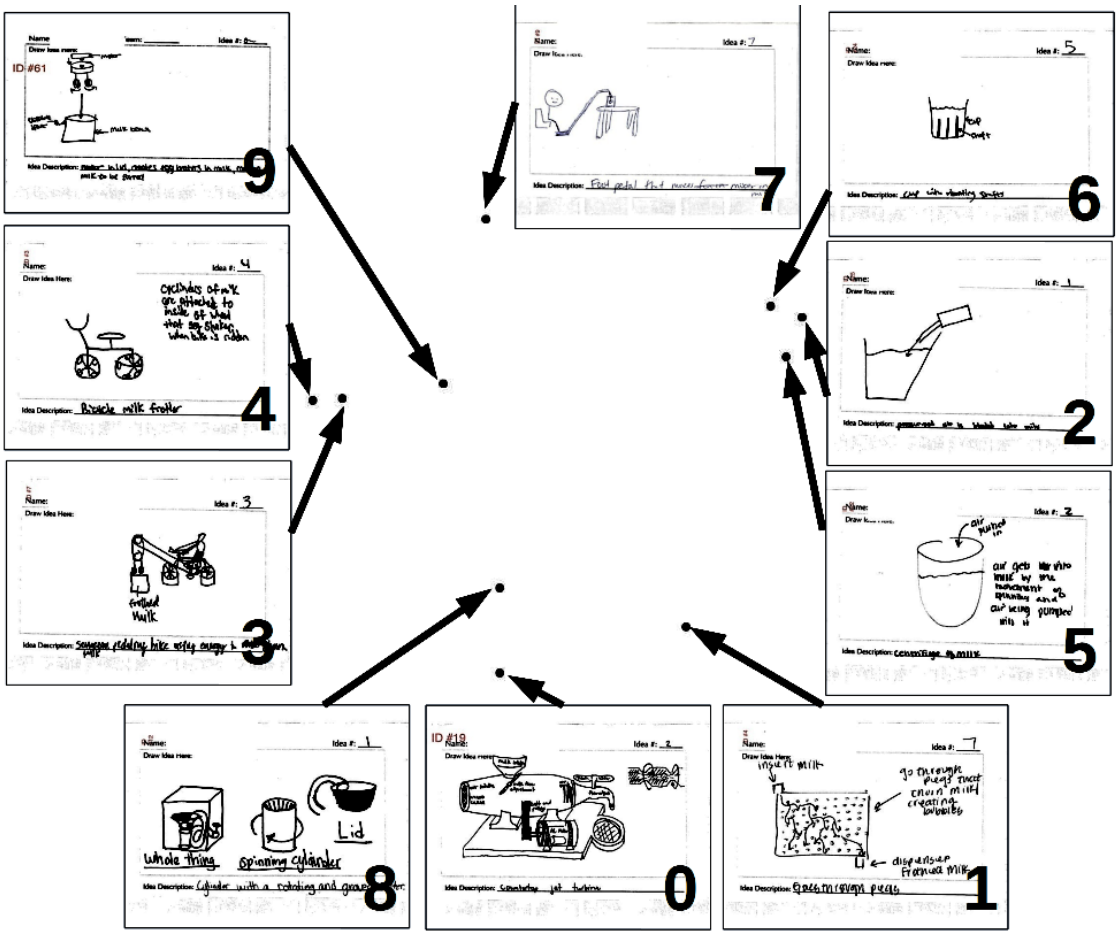

FIGURE 9. Idea map obtained by combining triplets from all raters. Id of each sketch is at bottom right corner.

of raters.

Sketches that are difficult to judge Different sketches have different levels of complexity. Some sketches in a triplet query can be considered similar/dissimilar based on multiple factors due to their design complexity (like sketch 0) but others may be simple in design and judged on fewer factors (like sketch 2). Finding sketches that are consistently difficult to judge by raters is important, as it can help understand features within these difficult sketches which cause disagreement among raters. To understand which sketches are more ambiguous or are difficult to rate, we measure the total number of times a sketch appears in triplets where raters disagreed. For instance, if 50% of raters give Design B as triplet response and other 50% give Design C, then all three sketches in this triplet are considered difficult to rate. We measure disagreement by the Shannon entropy of all responses and we calculate the score of each sketch by adding the entropy from all triplets for all raters in which it appears. Using this score, we find that sketch 8 has the highest disagreement score among raters, followed by sketch 0. Sketch 1 followed by sketch 6 have least disagreement scores. This indicates whenever sketch 8 appeared in a triplet, raters were more likely to give different responses. One possible reason for this can be design complexity. Sketch 8 and sketch 0 have many moving parts and are more detailed sketches, hence they can be interpreted differently by different raters compared to some other sketches which are simpler in design.

In the next section, we show that the embedding obtained by combining the triplets of multiple raters is quite robust. We show this using two experiments. First, we reduce the number of triplets available to derive the embedding and show that we can obtain a similar map using only a small fraction of triplet ratings originally used. Second, we add noise to the triplet ratings by flipping a percentage of triplets (simulating mistakes by raters) and show that these maps are resilient to significant levels of noise too.

Rater ID	Self Consistency	Top 3 Most Novel Sketch	# Transitive Violations	Percentage not satisfied
1	91.6%	5, 2, 4	5	17
2	50.0%	6, 0, 2	5	21
3	83.3%	1, 2, 7	5	20
4	75.0%	4, 0, 6	10	20
5	75.0%	2, 8, 5	10	21
6	58.3%	1, 4, 5	20	27
7	41.6%	4, 1, 2	8	15
8	41.6%	1, 7, 4	20	26
9	58.3%	0, 6, 1	11	16
10	75.0%	4, 0, 1	12	19
11	58.3.0%	5, 6, 2	5	16

TABLE 2. Rater performance and top three novel items for different raters of experiment 2 on design sketches

.

Number of triplets needed

As mentioned before, we had collected 360 similarity judgments from 11 raters for experiment 2. This task is time consuming and difficult to scale as the number of sketches grow. However, past researchers have found that one can obtain a meaningful embedding with fewer triplets than all triplets [37]. To empirically estimate how many triplets are needed to obtain an embedding close to the one obtained in Fig. 9, we varied the number of triplet ratings available to us and found different embeddings. As different embeddings cannot be directly compared, we calculate the euclidean distance matrix for a baseline embedding (that is distance between each sketch with other sketch). Then for any new embedding, we calculated mean squared error between new distance matrix and the baseline embedding. For any given percentage of triplets, we performed 100 runs with different subsets. Figure 10 shows the resultant median mean squared difference along with 5^{th} and 95^{th} percentile. We found that using a small fraction of, say, 40% of available triplets, the median error was only 0.016. With such low error, the resultant map obtained seemed quite similar to Fig. 9. Hence, we empirically verified that one can significantly reduce the number of triplets needed to find similar embeddings. This allows us to investigate approaches to reduce queries and make the process more scalable. In future work, we will investigate active learning approaches to minimize the number of triplet queries needed to construct meaningful embeddings.

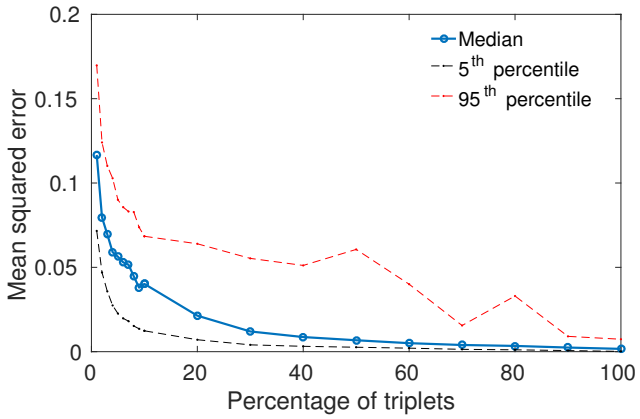

FIGURE 10. Mean squared error between distance matrices of embedding shown in Fig. 9 and embedding obtained using fewer triplets. 100 runs with different subsets used to obtain embeddings.

Effect of noise

In Table 2, we noticed that a few raters have low self consistency and have multiple transitive violations. How does this noise affect their idea map? We used all of the 3960 triplets obtained from 11 raters, but randomly flipped the response for a percentage of those triplets. This situation can occur in cases where rater accuracy goes down due to fatigue, when a few raters intentionally lie about similarity judgments or due to human error. Figure 11 shows the variation of the mean squared error from the baseline idea map with increasing noise percentage. When 25% of triplets are flipped, the median mean squared error is still only 0.018 (implying that there is little change in map). This shows that although increasing noise changes the idea map, this approach is still resilient to significant levels of noise.

DESIGN IMPLICATIONS

In this paper, we have provided preliminary results to visualize design ideas on a map and measure their novelty. Our experimental results have wide ranging implications in many design applications as listed below:

1. Generating idea maps using triplet queries is not limited to sketches and can be used for other type of design artifacts like CAD models or text documents to assess human perceived similarity. For larger datasets, one can use a small sample of design ideas with triplet queries to understand features which are given more importance in defining similarity of ideas. These features can then be used to build feature trees for the entire dataset.
2. Generating such maps can help in understanding the design domain. For instance, one can use maps to understand what features are more important in defining similarity between ideas. We find in our experimental results that raters form

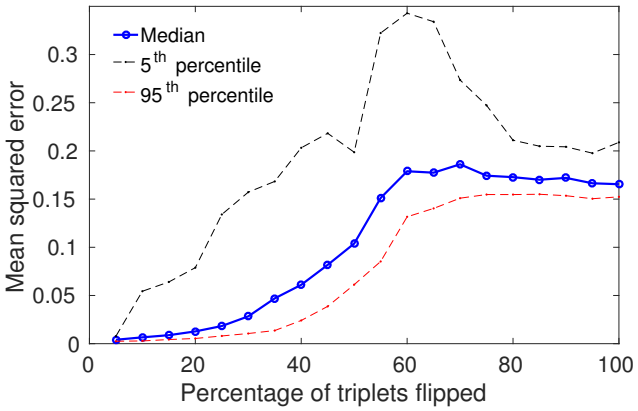

FIGURE 11. Mean squared error between distance matrices of embedding shown in Fig. 9 and embedding obtained using noisy triplets. We perform 100 runs with different subsets to obtain the embeddings. This shows that even when raters give small percentage of false ratings, idea maps are robust.

identifiable clusters in idea maps. This could mean a whole new way of finding and studying fine-grained details in how people reason about concepts and designs. One can also measure changes in idea maps of a person or team before and after some trigger event (like showing analogies) to understand change in perception of design space.

3. In our experimental results, we found that humans, even experts, are surprisingly inconsistent. This measure of inconsistency provides some evidence that subjective novelty ratings may often be inaccurate. Our experiments throw light on the observation that if human raters are inconsistent in comparing similarity of sets of three ideas, then how can we be sure that this inconsistency does not translate when they provide subjective novelty ratings? The latter task essentially requires comparing an idea with all other ideas in the domain, which is strictly harder problem than comparing three items at a time.
4. As raters are often inconsistent in their responses, we also show that triplet embeddings are fairly robust and can handle large noise conditions. This makes our method well suited for many applications where ratings are noisy or ambiguous. In comparing different novelty measurement methods, we believe that future studies should take into account robustness to noise too.
5. As shown in clustering of raters, we can measure similarity between raters from their triplet responses. This similarity measure can be used to find groups of similar raters. These groupings can be used to find aggregated maps for different groups and study differences between idea maps of a group of raters. For example, it can help to unpack differences in how experts rate items compared to novices. Measuring differences between raters can help in training them too, by understanding what features someone isn't paying attention to and providing appropriate intervention to increase inter-rater reliability. By following our study with qualitative questions, one can also understand how individuals came up with criteria to decide between triplets.
6. We provide a principled way of finding hard-to-judge concepts/designs. Finding these designs is important when assembling ground sets for things like verifying new metrics or the correct implementation of existing one. One can also allocate experts to rate hard-to-judge designs and use novices for easier designs.
7. Finally, finding accurate similarity representation paves the way of defining new family of variety and novelty metrics, which can help in assessment of ideas. In this paper, we have used sum of distances on a map as a measure of novelty, but other measures can also be defined to quantitatively measure novelty. For instance, after obtaining an embedding, one can use kernel PCA [38] to estimate novelty. One can also use volume based coverage methods like Determinantal Point Processes (DPP) [39] to give high score to ideas which have highest marginal gain in coverage.

LIMITATIONS

However, before adopting this methodology, one should be aware of various assumptions and limitations. Here we list few main limitations and future work directions to address them. Firstly, we have used two small datasets of ten items to demonstrate our results. The number of triplets required for a complete ordering is proportional to cube of the number of design items. This makes application to large datasets seem difficult. However, we show in our experimental studies that complete triplet set may not be needed to obtain meaningful embedding. In future work, we plan to use active learning to reduce the number of queries and study idea maps for larger datasets.

Secondly, the non-metric nature of queries creates few issues. First, it is insufficient to simply aim to satisfy the triplet constraints in the embedding through pairwise distances. It is possible to construct very different embeddings whilst satisfying the same percentage of the similarity triplets. This allows us to use further information from users to select between different possible embeddings. In future work, we aim to further optimize the idea map by incorporating additional constraints based on queries from raters. Apart from having multiple possible embeddings, measuring novelty on a low dimensional embedding obtained using metric distances is not straightforward due to non-metric nature of queries. Summing the euclidean distances on the 2-d idea map assumes metric space. To address this problem, we aim to investigate non-metric novelty measures in our future work.

Thirdly, we assume that design sketches exist on a 2-D em-

bedding and novelty can be interpreted as distance from all other items on this embedding. The 2-D assumption is important for map interpretability but may not be true for some design domains. There is also potential to extend the formulation of novelty we used. While current metric is simple and straightforward, it may have some unexpected limitations when designs are clustered. In future work, we aim to compare and contrast different ways to obtain maps and to measure novelty of items once the map is obtained.

Finally, different raters may use different criteria in deciding whether Item A is more similar to Item B or Item C. Ideas were only assessed by the raters at the idea level, not the feature level. Although, averaging the results of multiple raters provides a good estimate of aggregate view, the problem is inherently of multiple views. In future work, we will explore directly optimizing for multiple maps using multi-view triplet embeddings [40]. This will allow us to obtain multiple maps for each rater corresponding to different factors they considered.

CONCLUSION

In this paper, we proposed a method to find idea maps or two dimensional embedding of design ideas using triplet comparisons. We showed how these idea maps can be used to explain and measure novelty of ideas using two domains as examples—a set of polygons with known differentiation factors and a set of milk frother sketchers whose factors are unknown. These maps also highlighted interesting properties of how raters chose to differentiate concepts and how to group raters by similarity. In future work, we aim at three main extensions. First, by using active learning, we aim to extend this method to larger datasets with fewer triplet queries. Second, we aim to compare maps generated by triplets with map directly generated by a person on how they think sketches should be placed. We will also compare maps of different groups of people like experts vs novices to understand how people reason about ideas. Finally, we aim to optimize idea maps using additional information from raters and use multi-view triplet embeddings to study different attributes which may affect similarity of ideas.

ACKNOWLEDGEMENT

This material is based upon work supported by the National Science Foundation under Grant No. 1728086.

REFERENCES

[1] Starkey, E., Toh, C. A., and Miller, S. R., 2016. "Abandoning creativity: The evolution of creative ideas in engineering design course projects". *Design Studies,* ***47***, pp. 47–72.

[2] Hammedi, W., van Riel, A. C., and Sasovova, Z., 2011. "Antecedents and consequences of reflexivity in new product idea screening". *Journal of Product Innovation Management,* ***28***(5), pp. 662–679.

[3] Lopez-Mesa, B., and Vidal, R., 2006. "Novelty metrics in engineering design experiments". In DS 36: Proceedings DESIGN 2006, the 9th International Design Conference, Dubrovnik, Croatia.

[4] Sarkar, P., and Chakrabarti, A., 2011. "Assessing design creativity". *Design Studies,* ***32***(4), pp. 348–383.

[5] Johnson, T. A., Cheeley, A., Caldwell, B. W., and Green, M. G., 2016. "Comparison and extension of novelty metrics for problem-solving tasks". In ASME 2016 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference, American Society of Mechanical Engineers, pp. V007T06A012–V007T06A012.

[6] Verhaegen, P.-A., Vandevenne, D., and Duflou, J., 2012. "Originality and novelty: a different universe". In DS 70: Proceedings of DESIGN 2012, the 12th International Design Conference, Dubrovnik, Croatia.

[7] Maher, M. L., and Fisher, D. H., 2012. "Using ai to evaluate creative designs". In DS 73-1 Proceedings of the 2nd International Conference on Design Creativity Volume 1.

[8] Chen, L., Xu, P., and Liu, D., 2016. "Experts versus the crowd: a comparison of selection mechanisms in crowdsourcing contests".

[9] Chen, L., and Liu, D., 2012. *Comparing strategies for winning expert-rated and crowd-rated crowdsourcing contests: First findings*, Vol. 1. 12, pp. 97–107.

[10] Green, M., Seepersad, C. C., and Hölttä-Otto, K., 2014. "Crowd-sourcing the evaluation of creativity in conceptual design: A pilot study". In ASME 2014 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference, American Society of Mechanical Engineers, pp. V007T07A016–V007T07A016.

[11] Surowiecki, J., 2004. "The wisdom of crowds: Why the many are smarter than the few and how collective wisdom shapes business". *Economies, Societies and Nations,* ***296***.

[12] Hennessey, B. A., and Amabile, T. M., 1999. "Consensual assessment". *Encyclopedia of creativity,* ***1***, pp. 347–359.

[13] Licuanan, B. F., Dailey, L. R., and Mumford, M. D., 2007. "Idea evaluation: Error in evaluating highly original ideas". *The Journal of Creative Behavior,* ***41***(1), pp. 1–27.

[14] Shah, J. J., Kulkarni, S. V., and Vargas-Hernandez, N., 2000. "Evaluation of idea generation methods for conceptual design: effectiveness metrics and design of experiments". *Journal of Mechanical Design,* ***122***(4), pp. 377–384.

[15] Verhaegen, P.-A., Vandevenne, D., Peeters, J., and Duflou, J. R., 2013. "Refinements to the variety metric for idea evaluation". *Design Studies,* ***34***(2), pp. 243–263.

[16] Baer, J., 2012. "Domain specificity and the limits of cre-

ativity theory". *The Journal of Creative Behavior,* ***46***(1), pp. 16–29.
[17] Casakin, H., and Kreitler, S., 2005. "The nature of creativity in design". *Studying Designers,* ***5***, pp. 87–100.
[18] Brown, D. C., 2014. "Problems with the calculation of novelty metrics". In Proc. Design Creativity Workshop, 6th Int. Conf. on Design Computing and Cognition (DCC14).
[19] Richardson, T., Nekolny, B., Holub, J., and Winer, E. H., 2014. "Visualizing design spaces using two-dimensional contextual self-organizing maps". *AIAA Journal,* ***52***(4), pp. 725–738.
[20] Tang, J., Liu, J., Zhang, M., and Mei, Q., 2016. "Visualizing large-scale and high-dimensional data". In Proceedings of the 25th International Conference on World Wide Web, International World Wide Web Conferences Steering Committee, pp. 287–297.
[21] Maaten, L. v. d., and Hinton, G., 2008. "Visualizing data using t-sne". *Journal of machine learning research,* ***9***(Nov), pp. 2579–2605.
[22] Chen, W., Fuge, M., and Chazan, J., 2017. "Design manifolds capture the intrinsic complexity and dimension of design spaces". *Journal of Mechanical Design,* ***139***(5), p. 051102.
[23] Li, L., Malave, V., Song, A., and Yu, A. J., 2016. "Extracting human face similarity judgments: Pairs or triplets?". *Journal of Vision,* ***16***(12), pp. 719–719.
[24] Torgerson, W. S., 1958. "Theory and methods of scaling.".
[25] van der Maaten, L., and Weinberger, K., 2012. "Stochastic triplet embedding". In 2012 IEEE International Workshop on Machine Learning for Signal Processing, pp. 1–6.
[26] Stewart, N., Brown, G. D., and Chater, N., 2005. "Absolute identification by relative judgment.". *Psychological review,* ***112***(4), p. 881.
[27] Agarwal, S., Wills, J., Cayton, L., Lanckriet, G., Kriegman, D., and Belongie, S., 2007. "Generalized non-metric multidimensional scaling". In Artificial Intelligence and Statistics, pp. 11–18.
[28] Tamuz, O., Liu, C., Belongie, S., Shamir, O., and Kalai, A. T., 2011. "Adaptively learning the crowd kernel". In Proceedings of the 28th International Conference on International Conference on Machine Learning, ICML'11, Omnipress, pp. 673–680.
[29] Sankaranarayanan, S., Alavi, A., and Chellappa, R., 2016. "Triplet similarity embedding for face verification". *arXiv preprint arXiv:1602.03418*.
[30] Nhat, V. D. M., Vo, D., Challa, S., and Lee, S., 2008. "Non-metric mds for sensor localization". In Wireless Pervasive Computing, 2008. ISWPC 2008. 3rd International Symposium on, IEEE, pp. 396–400.
[31] Haghiri, S., Ghoshdastidar, D., and von Luxburg, U., 2017. "Comparison-based nearest neighbor search". In Artificial Intelligence and Statistics, pp. 851–859.
[32] Ukkonen, A., Derakhshan, B., and Heikinheimo, H., 2015. "Crowdsourced nonparametric density estimation using relative distances". In Third AAAI Conference on Human Computation and Crowdsourcing.
[33] Demiralp, Ç., Bernstein, M. S., and Heer, J., 2014. "Learning perceptual kernels for visualization design". *IEEE transactions on visualization and computer graphics,* ***20***(12), pp. 1933–1942.
[34] Siangliulue, P., Arnold, K. C., Gajos, K. Z., and Dow, S. P., 2015. "Toward collaborative ideation at scale: Leveraging ideas from others to generate more creative and diverse ideas". In Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing, ACM, pp. 937–945.
[35] Lin, H., and Bilmes, J., 2011. "A class of submodular functions for document summarization". In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1, Association for Computational Linguistics, pp. 510–520.
[36] Toh, C. A., and Miller, S. R., 2016. "Choosing creativity: the role of individual risk and ambiguity aversion on creative concept selection in engineering design". *Research in Engineering Design,* ***27***(3), pp. 195–219.
[37] Amid, E., Vlassis, N., and Warmuth, M. K., 2016. "Low-dimensional data embedding via robust ranking". *arXiv preprint arXiv:1611.09957*.
[38] Hoffmann, H., 2007. "Kernel pca for novelty detection". *Pattern Recognition,* ***40***(3), pp. 863–874.
[39] Ahmed, F., and Fuge, M., 2018. "Ranking ideas for diversity and quality". *Journal of Mechanical Design,* ***140***(1), p. 011101.
[40] Amid, E., and Ukkonen, A., 2015. "Multiview triplet embedding: Learning attributes in multiple maps". In International Conference on Machine Learning, pp. 1472–1480.