Unpacking subjective creativity ratings: Using triplet queries to find idea maps

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

Sharath Kumar Ramachandran
School of Engineering Design, Technology
and Professional Programs
The Pennsylvania State University
University Park, PA
Email: sharath@psu.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 design domain. We also propose that novel ideas can be identified by finding outliers on these idea maps. To demonstrate our method, we conduct experimental evaluations on two datasets of colored polygons (known answer) and milk frother sketches (unknown answer). We show that these maps shed light on factors considered by raters in judging idea similarity. We also show robustness of idea maps to fewer ratings or noisy ratings. We compare idea maps generated using triplet comparisons to physical maps manually made by the human subjects. This method provides a new direction of research into deriving ground truth novelty metrics by combining human judgments and computational methods.

1 Introduction

Creativity is the driving force of innovation in the design industry. Despite many methods to help designers enhance novelty of generated ideas, not much research has focused on what happens after this generation [1]. One of the main problems that design managers face after an ideation exercise is how to judge submitted ideas. 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

[*] Address all correspondence to this author.

management [2]. However, many of the existing methods of idea evaluation are inherently subjective. An emerging thread of research within idea evaluation is on attempts to quantitatively assess creativity of ideas [3–5]. Creativity of ideas is often viewed as the comparison of design ideas for quality and novelty. Quality is a measure of the designs' performance [6] and can be defined using multiple domain dependent factors like functionality, feasibility, usefulness, impact, investment potential, scalability, *etc.* In contrast to quality, novelty represents the uniqueness of an idea or how different it is from other designs in its class [7]. Both these metrics can be measured by human judges or using automated methods.

When ideas are being judged by humans, the judges can be 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. As an alternate to expensive experts, crowds have been proposed [10] to evaluate ideas due to their large diversity of viewpoints, knowledge and skills [11]. However, there is no clear evidence demonstrating that crowds can be used as a proxy for experts' evaluations to assess a large number of ideas [12]. In this paper, we focus on uncovering factors in subjective novelty ratings of people and do not differentiate between experts, non-experts or crowd voters.

Whether ideas are judged on novelty by crowds or experts, there are two important research issues that are key to idea evaluation— "What scale is used by people to judge novelty of ideas?" and "How can one explain the decision making process?" 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 a two dimensional Euclidean space—for raters and then estimating the novelty of those ideas from those maps.

2 Related Work

In this section, we review research related to creativity ratings and design space visualization, which relate directly to our work. The key contribution of our work is combining methods from creativity ratings and design space visualization to help create explainable metrics.

2.1 Creativity Ratings

In the social sciences, creativity is often measured subjectively through the Consensual Assessment Technique (CAT) [13]. They define a creative idea as something that experts in the idea's or project's focus area independently agree is creative. CAT is considered one of the gold standards for creativity assessment as it can reliably assess creativity, through the consensual assessments of domain experts. 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 design, experts may not remember all existing designs similar to it or they underestimate the originality of truly novel ideas [14]. By using different attributes or different criteria of evaluation within the same attribute, it is possible that experts will 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, 15, 16]. For example, one of the commonly used tree-based metrics [15] breaks down creativity into quantity, quality, novelty, and variety. These methods are widely adopted in engineering due to limited rater bias [17]. The resultant novelty score of an idea depends on which attributes are considered in the tree and may vary between two different raters or trees [18]. Despite the 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 [19, 20]. In this paper, we propose a third approach, which combines the strengths of both methods by asking simple subjective queries from raters and then using computational methods to estimate idea novelty. In the process, we also generate idea maps, which are used to visualize the design domain. Next, we discuss the efforts in design space visualization to give insights to designers about their domain.

2.2 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 [21] 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 [22, 23]. 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 people about how similar ideas are.

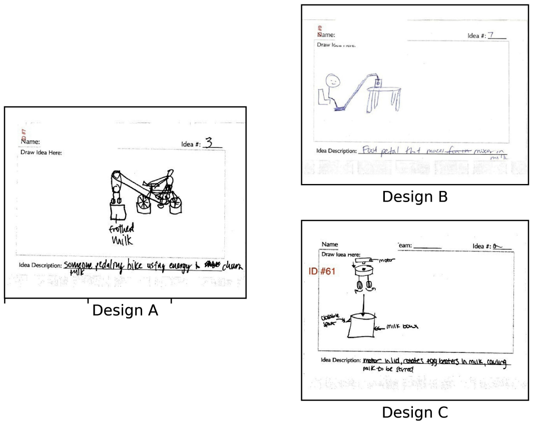

Fig. 1. Example of triplet query asked from raters in our experiment. Rater answers the question: "Which design is more similar to design A?"

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) [24]. Methods like classical multi-dimensional scaling [25] 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 [26].

As humans are better at comparing items than giving absolute scores [27], 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) [28], Crowd Kernel Learning [29] and Stochastic Triplet Embedding [26]. Such methods take in triplet ratings and output either an embedding or similarity kernel between items which best satisfy human triplet responses.

Techniques for capturing similarity among items using triplets have been applied in many areas like computer vision [30], sensor localization [31], nearest neighbor search [32] and density estimation [33]. In [34], 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. Siangliulue *et al.* [35] 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 as we use idea maps to measure novelty of design ideas and try to uncover attributes behind the decision making of raters.

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 discussion on extension of this method to study design novelty.

3 Methodology

Below, we discuss how triplet responses can be used to estimate idea maps and define two novelty metrics based on these maps.

3.1 Idea Map Generation

Given a set of N designs, we first generate all possible triplet queries from them. This set of queries is given to raters as surveys. After collecting responses, we use the Generalized Non-metric Multidimensional Scaling (GNMDS) technique [28] to find embeddings of design ideas.[1] The idea map obtained by applying GNMDS to the triplet responses by a rater tries to satisfy a majority of the triplets. To do so, GNMDS finds a low-rank kernel matrix K in such a way that the pairwise distances

[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.

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 query with three design sketches used in our study. We represent the response of the rater to any query as ‘ABC’ or ‘ACB’. Response coded as ‘ABC’ means Design A is closer to Design B than Design C and response coded as ‘ACB’ means Design A is closer to Design C than Design B. GNMDS method 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.

3.2 Measuring Novelty on a Map

Given an idea map, our goal is to calculate novelty score of each idea. 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 novel within the set. As novelty of an item in a set can be interpreted as how unique or dissimilar an item is [7], the problem is equivalent to finding ideas which are distant from all other ideas on the map. However, many different methods exist to find outliers on a two dimensional map. Here, we define two such metrics which give a high score to ideas which are away from everyone else on a map. We name these metrics as Nov_{sum} and Nov_{cent}, which score any item i as follows:

$$Nov_{sum}\ (i) = \sum_{j=1}^{N} \sqrt{(X_i - X_j)^2 + (Y_i - Y_j)^2} \tag{1}$$

$$Nov_{cent}\ (i) = \sqrt{(X_i - X_c)^2 + (Y_i - Y_c)^2} \tag{2}$$

Here X_i, Y_i are the 2-D coordinates of idea i. X_c is the 2-D coordinates of the centroid of all ideas. Nov_{sum} defines novelty of an idea in a set as the sum of distances from the idea to all other ideas. This simple formulation has been used in the past for document summarization to define representative items [36]. Nov_{cent} defines the novelty of an idea as the distance from the centroid and has been used in [6] to measure novelty. The centroid is a theoretical point in the space, created by averaging the attributed values across all designs in the space. By giving high score to ideas furthest away, both metrics allow us to rank order all ideas by their novelty.

We experimented with a few other methods to measure novelty of items on a map but chose these two metrics as they are a) easy to compute and b) make few assumptions about the distribution of ideas or how ideas are clustered in the map. Note that it is possible to discuss many more metrics for novelty detection on a two-dimensional map, however, we only use these two metrics to show how triplet embeddings enable novelty calculation. Finding the best novelty metric (depending on how one defines best metric and if such metric generalizes to all domains) for any given domain is outside the scope of this paper.

3.3 Measuring Rater Performance

Triplet responses given by raters can vary in accuracy or reliability due to factors like rater expertise or motivation. However, it is difficult to assess the quality of triplets by measuring intra-rater reliability, as they are subjective assessment of how a rater views the similarity of ideas. Instead, 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 triplet queries, which are repeats of existing queries. Second, we measure the number of violations a rater makes in the transitive property of inequality; for example, suppose a rater gives two responses as ABC and CAB, which means that she finds item A more similar to item B and item C more similar to item A. These responses imply $AB < AC$ and $CA < CB$, where AB is measure of similarity between idea A and idea B. These two inequalities imply the third inequality, that idea B is more similar to idea A ($BA < BC$). If this rater provides a third triplet response of BCA indicating idea B is 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 count the total number of transitive 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 these metrics but this information could be incorporated in future studies to give more importance to idea maps of raters who are more self-consistent.

3.4 Measuring Map Similarity

To find similarity between two idea maps, we employ three different methods by comparing: a) 2-D positions, b) Distance vectors, c) Overlap between triplets obtained from each map.

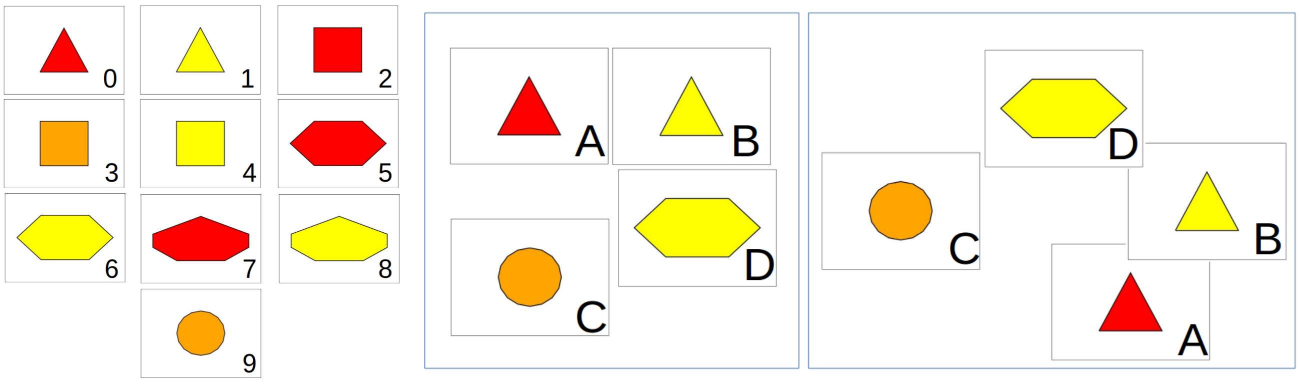

Fig. 2. a) Dataset of ten polygons used in experiment 1 b) Two idea maps with 4 items each. Although these maps look different, they satisfy the same set of triplet queries.

To compare the 2-D position of points on two maps, they should be on the same scale. However, maps obtained by triplets or drawn by people can be on different scales and maybe rotated or translated. To overcome this problem, we use procrustes analysis [37] to find the optimal scaling/dilation, rotations, and reflections such that the sum of the squares of the pointwise differences between the two input datasets is minimized. We call the least squared error after transformation of one map as the 'Disparity' score between the two sets of points.[2] However, this measure is dependent on an intermediary step of correctly solving another optimization problem, which may introduce error (if the procrustes transformation converges to a local minima).

To get more confidence in comparing two maps, we define two more map similarity methods. In the distance based method, we calculate the euclidean distance vector of each point with every other point. For 10 points, we get 45 unique distances. We find the mean squared error (MSE) between the distance vectors of the two maps. Distances are rotation and translation invariant. We divide each distance vector by the maximum distance of that vector to make them scale invariant too. This resolves the issue of different map scaling by bounding the maximum distance for each map to one unit.

The above distance method gives a measure of how metric distances between two maps differ. However, as the maps are generated using non-metric triplets, maps with different spatial arrangements can still satisfy the same set of triplets. Hence, we propose a new non-metric similarity measure between two maps called "Triplet error". In this method, we generate a set of triplet responses corresponding to each map such that it satisfies the given map exactly. Let us call these sets S_1 and S_2. This set of triplet response can be different from the triplet set from which the map is generated (as we will see in our experimental results, maps may not satisfy a small proportion of triplets provided by raters). We count the number of triplet responses which are common between the two maps. Triplet error is defined as the percentage overlap between these two sets of triplets *i.e.* $\frac{|S_1 \cap S_2|}{|S_1|}$. Triplet error measures how the two maps compare in relative distances of items.

To explain triplet error, we take an example of comparing two maps with four items each as shown in Fig. 2 b). Visually the two maps look different. However, if we list triplet responses which satisfy the map on left side, we get the following set of twelve triplets: ABC, ABD, ACD, BAD, BAC, BDC, CAB, CDA, CDB, DBA, DCA and DBC. As mentioned before, item ABC means A is closer to B than C. If we list the triplets satisfying the map on the right side, we get the exact same set of triplet responses. Hence, the triplet error is zero between these two maps. This measure is independent of scaling and allows similar maps to have different spatial arrangements. In comparing maps in our experiment section, we report all three measures.

4 Experimental Results

To demonstrate our methodology, we consider two case studies. We chose 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. 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

[2] Score calculated using Python scipy library: `https://docs.scipy.org/doc/scipy-0.16.1/reference/generated/scipy.spatial.procrustes.html`

to generate meaningful idea maps. We also ask raters to generate physical maps directly and compare them to idea maps obtained using embeddings.

4.1 Experiment 1: Colored Polygons

Our dataset of ten polygons is shown in Fig. 2 a), 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 a) respectively. For this triplet, raters have to decide whether they find the red heptagon more similar to the yellow hexagon or the red square. One rater may prioritize color-based similarity to shape and thus answer "the red heptagon is more similar to the red square," while another may use closeness in area of polygons to answer "the red heptagon is more similar to the yellow hexagon". To gain insights into their decision making process, we also ask raters to explain their choice for 20 randomly selected triplets. These responses helped verify our hypotheses about the factors considered by each rater.

Automated rater To verify that the triplet generated maps correctly reflect provided triplet responses, we first use an automated rater who rates all triplet queries consistently based on a fixed set of rules. We define the rules such that this automated rater always rates polygons with the least difference in number of sides as more similar. When two polygons B and C have similar priority in previous rule, it selects the polygon which is more similar in color to base polygon A. 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, as expected. 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 are rated as more similar by the automated rater. Hence, this map can be considered a good representation of the triplet ratings provided by the automated 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.* We summarize our results for 11 raters in Table 1. Column 2 lists the self consistency score for each rater and column 5 lists the count of transitive violations. Column 3 and 4 provide the Top 3 items calculated using the two novelty metrics discussed before. Column 6 reports the percentage of triplet responses not satisfied by each map found using embedding method (lower is better).

Let us take the example of idea maps obtained for two raters (rater id 5 and rater id 9 from Table 1 respectively). Idea map of rater 5, shown in Fig. 4 places similar shaped polygons 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 to answer triplet queries. In contrast, the placement of similarly colored items together for map of rater 9 (Fig. 5) indicates that color is more important to her than shape. The orange square is closer to the orange circle in her map and far from similarly shaped squares.

When we look at the explanation provided by rater 5 for a subset of queries, 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 based on "color, shape, number of sides". Hence, the criteria used by individual raters are reflected in their idea maps, grouping similarly colored or shaped items together.

Given the idea maps of these ten polygons, one would expect the most novel item to be most dissimilar to all other polygons. For rater 5, 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.

Table 1 shows the top three most novel sketches for each rater using the two novelty metrics. We find that the two metrics give the same set of Top 3 items for 9 raters and remaining 4 have atleast 2 items common. This shows that the two metrics align in their novelty assessment. We also find that the orange circle (Polygon 9) appears in top three for most raters, indicating novelty metrics indicate 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 designed this dataset such that the circle is of a different color and unique shape compared to all other items in the set. The main takeaway from this experiment is that 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 individuals consider in deciding item similarity.

4.2 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 [1,38]. To create the original dataset, the authors recruited engineering

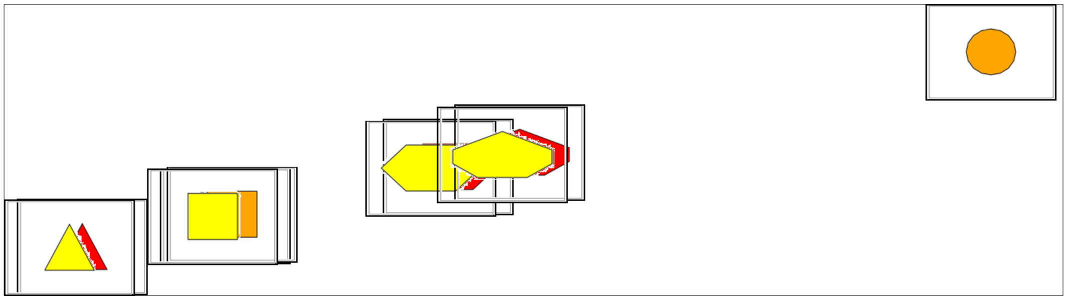

Fig. 3. Two dimensional embedding for automated rater for polygons

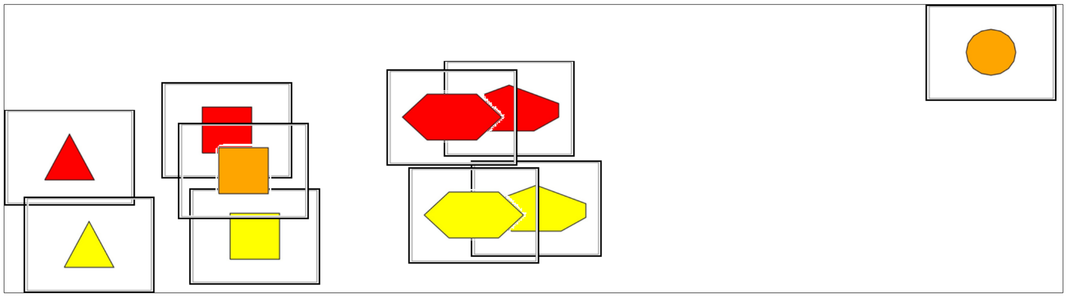

Fig. 4. Two dimensional embedding obtained from polygon dataset by rater 5, who uses number of sides as primary criteria for triplet decisions

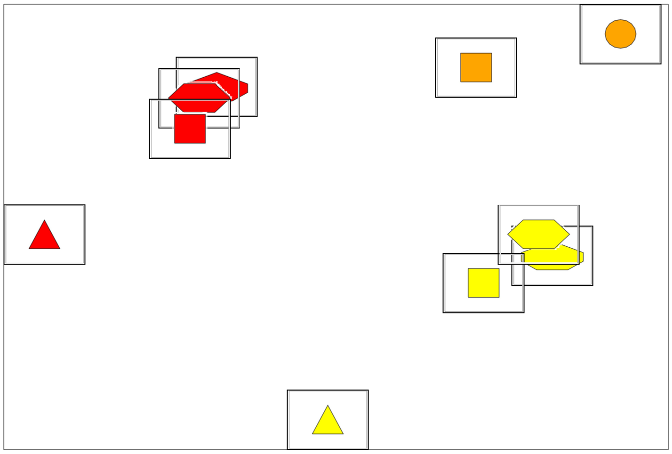

Fig. 5. Two dimensional embedding obtained from polygon dataset by rater 9, who uses 'color, shape, number of side' as criteria

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.[3]

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 a) and 7 b) 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 a) provides evidence that she might have grouped sketches which have cup to store milk in the design as more similar (as shown by

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

Rater id	Self consistency (%)	Top 3 Nov_{sum}	Top 3 Nov_{cent}	Transitive violations	Triplets not satisfied (%)
AR	100.0	9, 1, 0	9, 1, 0	0	5
1	83.3	9, 1, 0	9, 1, 0	2	11
2	100.0	9, 0, 8	9, 8, 0	3	11
3	83.3	9, 4, 8	9, 8, 4	3	15
4	75.0	9, 1, 0	9, 1, 8	2	15
5	100.0	9, 1, 0	9, 1, 0	0	1
6	100.0	9, 1, 0	9, 1, 0	0	15
7	91.6	9, 1, 0	9, 1, 0	0	15
8	91.6	9, 1, 6	9, 1, 6	8	21
9	83.3	1, 9, 0	1, 9, 0	9	22
10	83.3	9, 1, 3	9, 1, 0	4	10
11	100.0	9, 8, 1	9, 8, 6	0	15

Table 1. Rater performance and top three novel items for different raters of experiment 2 on design sketches
.

Rater id	Self consistency (%)	Top 3 Nov_{sum}	Top 3 Nov_{cent}	Transitive violations	Triplets not satisfied (%)
1	91.6	5, 2, 4	5, 2, 4	5	17
2	50.0	6, 0, 2	6, 0, 2	5	21
3	83.3	1, 2, 7	1, 7, 0	5	20
4	75.0	4, 0, 6	0, 4, 6	10	20
5	75.0	2, 8, 5	2, 8, 3	10	21
6	58.3	1, 4, 5	1, 4, 5	20	27
7	41.6	4, 1, 2	4, 2, 1	8	15
8	41.6	1, 7, 4	1, 4, 7	20	26
9	58.3	0, 6, 1	0, 6, 2	11	16
10	75.0	4, 0, 1	4, 0, 2	12	19
11	58.3%	5, 6, 2	5, 0, 6	5	16

Table 2. a) 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.

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 qualitatively analyzed their explanations. 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 are possibly grouped 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. Her top three most novel sketches were $0, 1$ & 6. Nov_{sum} metric finds sketches $4, 0$ & 1 as the top 3 ideas from her idea map while Nov_{cent} finds $4, 0$ & 2 as the top 3 items. While the rankings don't completely overlap, it should be noted that her top 3 sketches $(0, 1, 6)$ occur on the periphery of her idea map, showing that they are generally far away from other sketches. To further compare our results with existing methods, we also coded different attributes for all 10 sketches and use SVS metric to calculate their novelty scores. The scores are: 0.718, 0.585, 0.6, 0.692, 0.566, 0.483, 0.585, 0.612, 0.45 and 0.715 for sketches 1 to 10 respectively. Using SVS scores, we find Sketches $0, 9$ & 3 are most novel, which is different from the subjective ranking provided by the rater as well as the scores calculated using idea maps. The difference can be attributed to factors considered in SVS score calculation.

We also found differences between justifications given by expert 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 their maps to uncover factors considered by each group.

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

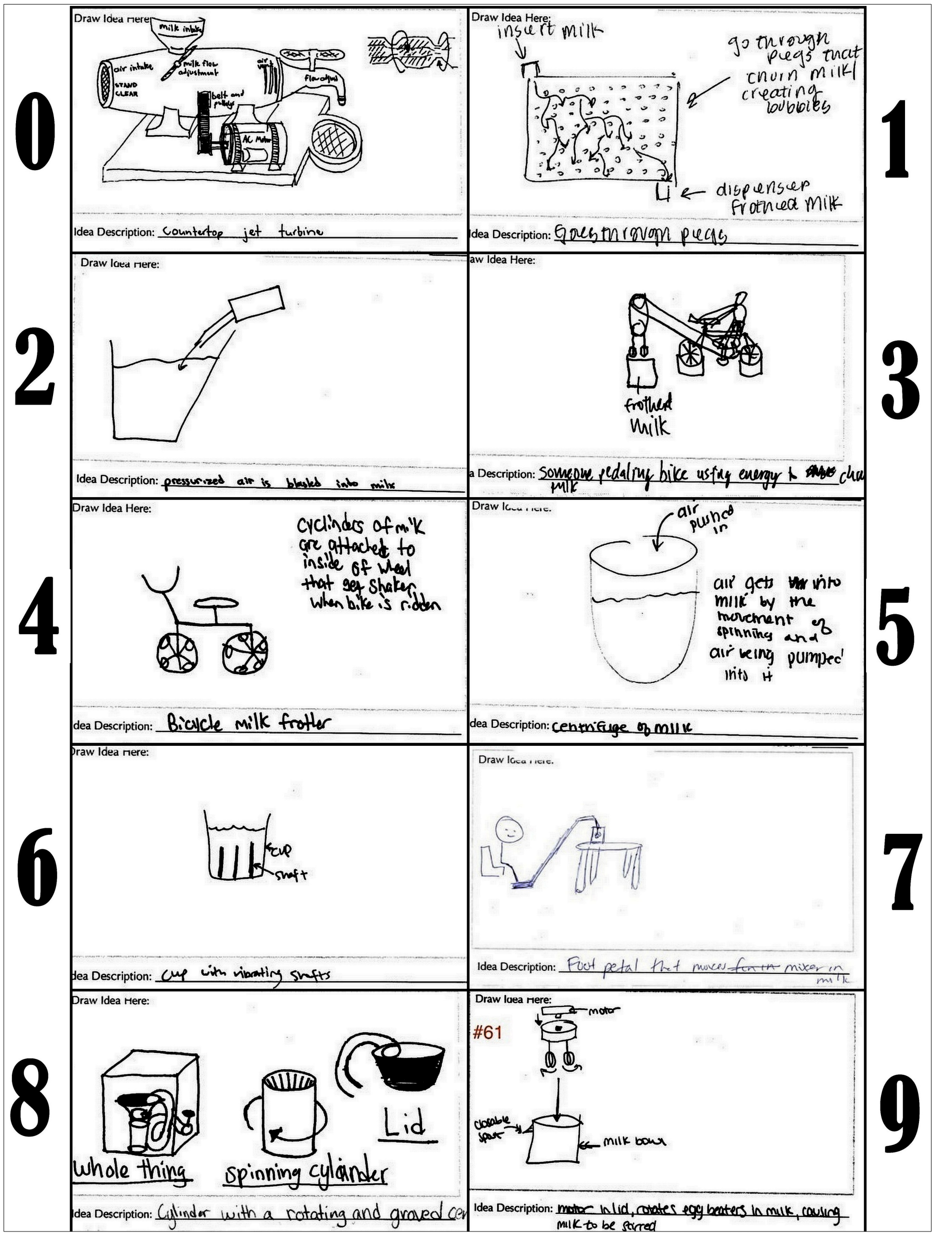
0
Draw Idea Here:
milk intake
air intake
milk flow adjustment
air vent
STAND CLEAR
belt and pulleys
AC Motor
Flow adjust
Idea Description: Countertop jet turbine
1
Draw Idea Here:
insert milk
go through pegs that churn milk creating bubbles
dispenser frothed milk
Idea Description: Goes through pegs
2
Draw Idea Here:
Idea Description: pressurized air is blasted into milk
3
Draw Idea Here:
frothed milk
Description: Someone pedaling bike using energy to ~~shake~~ churn milk
4
Draw Idea Here:
cyclinders of milk are attached to inside of wheel that get shaken when bike is ridden
Idea Description: Bicycle milk frother
5
Draw Idea Here:
air pushed in
air gets ~~the~~ into milk by the movement of spinning and air being pumped into it
Idea Description: centrifuge of milk
6
Draw Idea Here:
cup
shaft
Idea Description: cup with vibrating shafts
7
Draw Idea Here:
Idea Description: Foot petal that moves ~~forth~~ mixer in milk
8
Draw Idea Here:
whole thing
spinning cylinder
Lid
Idea Description: Cylinder with a rotating and grooved cen
9
Draw Idea Here:
#61
motor
closable spout
milk bowl
Idea Description: ~~motor~~ in lid, rotates egg beaters in milk, causing milk to be stirred

Fig. 6. Ten milk frother sketches used in experimental 2.

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, we combine the triplet responses of all raters and obtain a joint idea map. Fig. 8 shows the joint map of all 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) 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. 8, 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

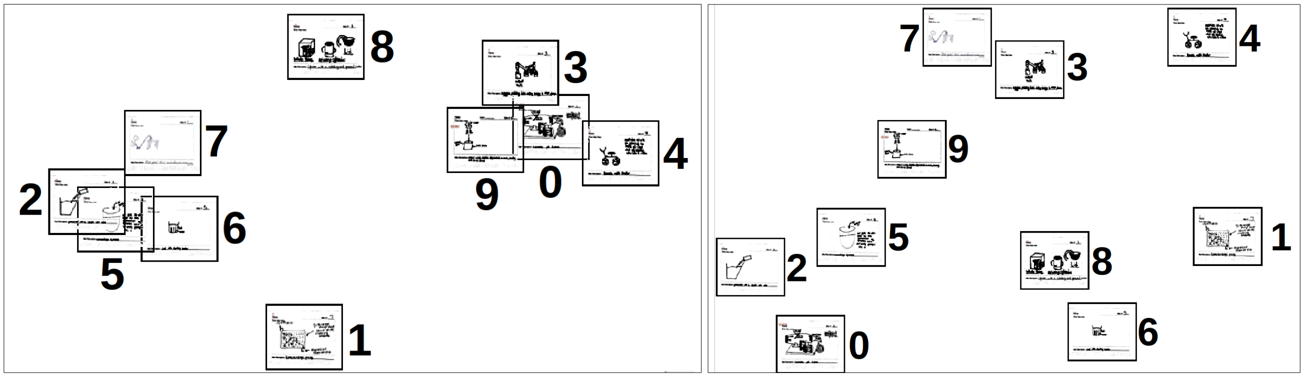

Fig. 7. a) Idea map of design sketches for rater 7. Center of the sketch represents the 2-D position of embedding. Two main clusters can be seen. b) Idea map of design sketches for rater 10. Center of the sketch represents the 2-D position of embedding.

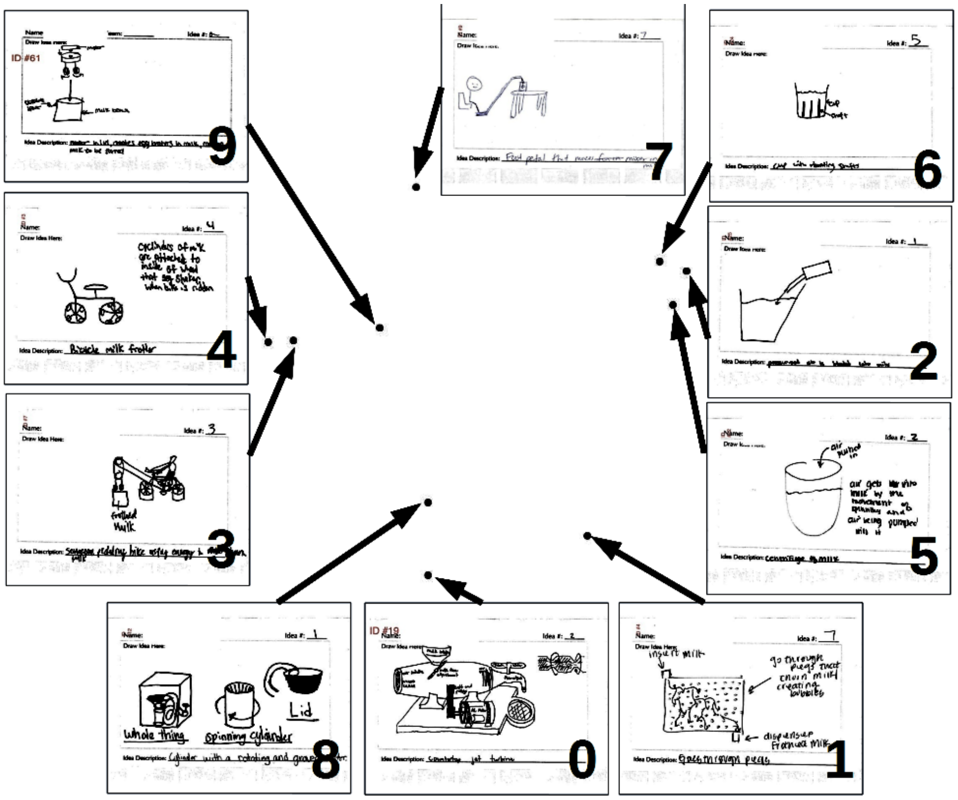

Fig. 8. Idea map obtained by combining triplets from all raters. Id of each sketch is at bottom right corner.

similarity. We have also shown how a joint map of multiple raters meaningfully groups sketches 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 multiple clustering methods to identify groups among these users and identified two clusters. We found 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

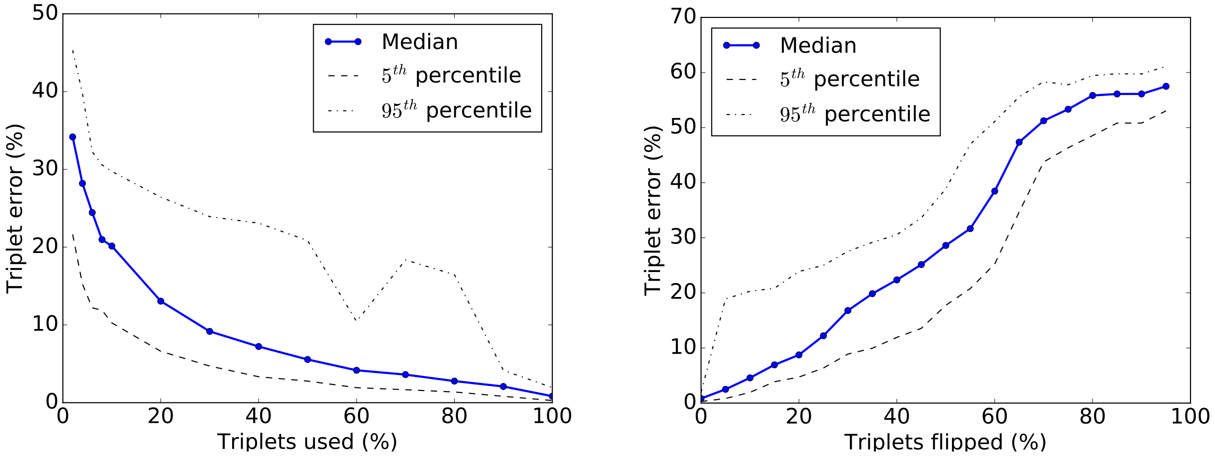

Fig. 9. a) Triplet error between idea maps of embedding shown in Fig. 8 and embedding obtained using a subset of triplet ratings. 100 runs with different subsets used to obtain embeddings. Using only 30% of triplets, median error is less than 10%. b) Triplet error between embedding generated using noisy triplets compared to embedding shown in Fig. 8. We perform 100 runs and flip a subset of triplets randomly to obtain the embeddings. Small increase in the median error shows that idea maps are robust to small percentage of false ratings by raters.

maps of group 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.

4.3 Maps using fewer triplets

As mentioned before, we collected 360 similarity judgments each from 11 raters for both experiments. 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 [39]. To empirically measure how many triplets are needed to obtain an embedding close to the one obtained in Fig. 8, we varied the number of triplet ratings available to us and found different embeddings. As different embeddings cannot be directly compared, we calculate the triplet error of each embedding with baseline embedding of Fig. 8. For any given percentage of triplets to be used, we performed 100 runs with different subsets. Figure 9 a) shows the resultant median triplet error along with 5^{th} and 95^{th} percentile. We found that using a small fraction of, say, 30% of available triplets, the median triplet error is only 9.1%. Hence, one can significantly reduce the number of triplets needed to find these embeddings. In future work, we will investigate active learning approaches to minimize the number of triplet queries needed to construct meaningful embeddings for larger datasets.

4.4 Maps using noisy triplets

In Table 2, we notice that a few raters have low self consistency scores and suffer from multiple transitive violations. To study how such noise can affect the idea map, we conduct an experiment to simulate noisy responses. We use all the 3960 triplet queries obtained from 11 raters, but randomly flip 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, rater changes the criteria to judge similarities while doing the survey, human error, *etc.* To measure the effect of noise, we assume the map shown in Fig. 9 b) as the ground truth and compare it to maps obtained from noisy labels using the triplet error metric. Figure 9 b) shows the variation of the triplet error from the baseline idea map (Fig. 8) with increasing noise percentage. When 25% of triplets are flipped, the median triplet error is only 8.3%. To understand how much triplet error is acceptable, we refer the readers to comparison of physical map with triplet map in next section. Here, triplet error of 18% can occur in reasonably similar maps with few items misaligned (Fig. 10). This shows that although increasing noise changes the idea map, this approach is still resilient to significant levels of noise.

5 Comparison with human generated maps

So far we have generated and compared idea maps created using only the triplet responses. How do these algorithmically-generated maps compare to a map that the same rater would generate directly (*i.e.*, by placing ideas on a 2D surface) without the intermediate step of answering triplets. In this section, we conduct additional experiment to generate idea maps directly from raters and then compare these idea maps with the maps generated using embedding methods.

5.1 Participants

Four subjects were selected from the group of raters that had participated in the triplet surveys. The subjects were selected based on their consistency in answering the triplet surveys. The participants comprised of 1 Faculty member, 2 Doctoral students and 1 Undergraduate student. The participants were given a six-month period between taking the survey and participating in the following experiment to avoid priming, and obtain results that are unbiased with respect to their original triplet responses.

5.2 Experimental setup

Each participant was provided with the same 10 idea sketches utilized in the triplet survey, printed on 8.5” x 5.5” sheets of paper. The order of the ideas was randomized for each participant. The subjects were required to pin the sketches on a 65” x 55” canvas, such that the distance between any two sketches would be proportional to how similar they were to each other. The sketches were allowed to overlap. The subjects were allowed to move the sketches multiple times, until they were satisfied with the idea map created. The participants were allotted a maximum time of 30 minutes for the activity. The participants were required to think aloud as they placed and moved the ideas around on the canvas. Throughout the activity, the participants were recorded using audio and video equipment. Figure 10 a) shows how the maps were pinned on a board by one of the raters participating in this experiment.

5.3 Comparison between automated and manual maps

We compare manual idea maps with automated idea maps (generated using triplet embedding methods), using the different metrics defined in the previous sections: 2-D position based, distance based and triplet error based.

Figure 10 b shows the manual map and automated map overlaid on each other for Rater 10. We notice that her automated and manual map align quite well, as seen by similar numbers (sketch id) positioned nearby each other. Table 3 summarizes the results for all four raters. Column 2 provides the percentage of triplets satisfied by the human-generated map. It measures percentage of triplet inequalities (from the survey taken by same rater) that are satisfied by the map generated by the person. Column 3 gives the triplet error between map obtained using automated method and manual map. We notice that this error increases for raters who have low self consistency (column 1). The disparity measure and mean squared error have a similar trend as triplet error. We notice that Rater 10 has the highest alignment between her manual and automated maps using all three metrics, while rater 5 has the least alignment. The lack of alignment can be explained by a variety of factors like change in similarity criteria or lack of difficulty in creating manual maps which satisfy all preferences.

6 Design implications and Limitations

In this paper, we propose using idea maps obtained from simple triplet queries to visualize design domain and measure idea 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

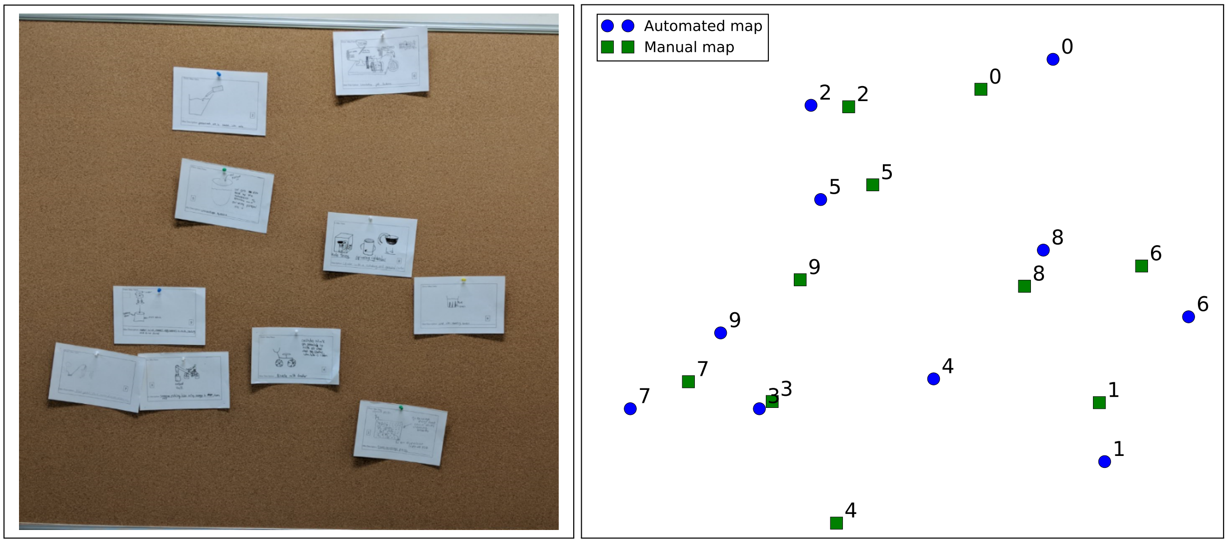

Fig. 10. a) Participant creating a map by positioning idea sketches on provided canvas. b) Correspondence between human generated map and triplet map for Rater 10 after Procrustes transformation. Apart from sketch 4, most sketches have minor relative displacements.

Rater id	% triplet satisfied	Triplet error (%)	Disparity	Distance MSE
1	26.9	28.1	0.31	0.064
3	34.1	36.2	0.47	0.082
5	39.1	41.7	0.52	0.092
10	25.8	18.1	0.15	0.023

Table 3. Comparison between maps drawn manually by four raters and their triplets and triplet embedding maps. We observe that manual maps are not great at satisfying triplets.

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 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 provide evidence that if human raters are inconsistent in comparing similarity of sets of three ideas, then this inconsistency may translate when they provide subjective novelty ratings too. 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 embedding methods and novelty metrics, future studies should take into account robustness to noise too.
5. As shown when clustering raters, we can measure the 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, or how groups of experts from different fields might differ. Measuring differences between raters can help in training them too, by understanding what features someone is not 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 come 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 as-

sembling 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 for defining new families of variety and novelty metrics, which can help assess ideas. In this paper, we have used simple novelty metrics like sum of distances on a map, but other measures can also be defined to quantitatively measure novelty. For instance, after obtaining an embedding, one can use kernel PCA [40] to estimate novelty. One can also use volume based coverage methods like Determinantal Point Processes (DPP) [41] to give high score to ideas which have highest marginal gain in coverage. Similarity representation for sketches allows us to use methods like diverse subset selection [42]— methods which traditionally need vector representation of design items.

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. In the naïve implementation, 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. 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 problems. It is insufficient to simply 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 as shown in Fig. 2 b). This allows us to use further information from users to select between different possible embeddings. Further research can be done in ways to optimize the idea map by incorporating additional user preference information. Apart from multiple possible embeddings, measuring novelty using metric distances is difficult due to non-metric nature of queries.

Thirdly, we assume that design sketches exist on a 2-D embedding 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 metrics are simple and straightforward, they may have some unexpected limitations when designs are clustered. In future work, we plan 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 [43]. This will allow us to obtain multiple maps for each rater corresponding to different factors they considered.

7 Conclusion

In this paper, we propose a method to find idea maps or two dimensional embedding of design ideas using triplet comparisons. We show how these idea maps can be used to explain and measure novelty of ideas. We use 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 highlight interesting properties of how raters chose to differentiate concepts and how to group raters by similarity. We compare our results using both completely automated method and using human generated maps. 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 code external human preferences into the optimization framework to find richer idea maps. Finally, we aim to extend this method to multi-view embeddings, to obtain idea maps corresponding to each factor considered by the rater and calculate feature specific novelty.

Acknowledgements

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] 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.

[7] 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.

[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] Görzen, T., and Kundisch, D., 2016. "Can the crowd substitute experts in evaluating creative jobs? the case of business models.". In ECIS, pp. Research–in.

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

[14] 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.

[15] 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.

[16] 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.

[17] Oman, S. K., Tumer, I. Y., Wood, K., and Seepersad, C., 2013. "A comparison of creativity and innovation metrics and sample validation through in-class design projects". *Research in Engineering Design,* **24**(1), pp. 65–92.

[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] Baer, J., 2012. "Domain specificity and the limits of creativity theory". *The Journal of Creative Behavior,* **46**(1), pp. 16–29.

[20] Casakin, H., and Kreitler, S., 2005. "The nature of creativity in design". *Studying Designers,* **5**, pp. 87–100.

[21] 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.

[22] 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.

[23] Maaten, L. v. d., and Hinton, G., 2008. "Visualizing data using t-sne". *Journal of machine learning research,* **9**(Nov), pp. 2579–2605.

[24] 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.

[25] Torgerson, W. S., 1958. "Theory and methods of scaling.".

[26] 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.

[27] Stewart, N., Brown, G. D., and Chater, N., 2005. "Absolute identification by relative judgment.". *Psychological review,* **112**(4), p. 881.

[28] 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.

[29] 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.

[30] Sankaranarayanan, S., Alavi, A., and Chellappa, R., 2016. "Triplet similarity embedding for face verification". *arXiv preprint arXiv:1602.03418*.

[31] Nhat, V. D. M., Vo, D., Challa, S., and Lee, S., 2008. "Nonmetric mds for sensor localization". In Wireless Pervasive Computing, 2008. ISWPC 2008. 3rd International Symposium on, IEEE, pp. 396–400.

[32] Haghiri, S., Ghoshdastidar, D., and von Luxburg, U., 2017. "Comparison-based nearest neighbor search". In Artificial Intelligence and Statistics, pp. 851–859.

[33] Ukkonen, A., Derakhshan, B., and Heikinheimo, H., 2015. "Crowdsourced nonparametric density estimation using relative distances". In Third AAAI Conference on Human Computation and Crowdsourcing.

[34] 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.

[35] 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.

[36] 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.

[37] Gower, J. C., 1975. "Generalized procrustes analysis". *Psychometrika,* **40**(1), pp. 33–51.

[38] 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.

[39] Amid, E., Vlassis, N., and Warmuth, M. K., 2016. "Low-dimensional data embedding via robust ranking". *arXiv preprint arXiv:1611.09957.*

[40] Hoffmann, H., 2007. "Kernel pca for novelty detection". *Pattern Recognition,* **40**(3), pp. 863–874.

[41] Ahmed, F., and Fuge, M., 2018. "Ranking ideas for diversity and quality". *Journal of Mechanical Design,* **140**(1), p. 011101.

[42] Ahmed, F., Fuge, M., and Gorbunov, L. D., 2016. "Discovering diverse, high quality design ideas from a large corpus". In ASME International Design Engineering Technical Conferences, ASME.

[43] Amid, E., and Ukkonen, A., 2015. "Multiview triplet embedding: Learning attributes in multiple maps". In International Conference on Machine Learning, pp. 1472–1480.